Multi-trait methods for genetic association testing

Porter, Heather Frances

Awarding institution:
King's College London

Multi-trait methods for genetic association testing

Heather Frances Porter

Thesis submitted for the degree of Doctor of Philosophy

MRC SGDP Centre, IoPPN, King's College London

September 2016

Abstract

The early stages of the genome-wide association study (GWAS) era were dominated by studies focusing on single phenotypes, while in recent years there has been growing interest in multi-trait GWAS. A wide variety of multi-trait GWAS methods have been developed, but publications introducing new methods are highly inconsistent in their evaluation of method performance, obscuring their relative merit. Facilitated by burgeoning national biobank resources, multi-trait analyses are set to become more routinely applied, making understanding their relative performance increasingly important.

We develop a simulation framework to model the complex networks underlying multivariate genetic epidemiology. We exploit our simulation framework to perform a comprehensive comparison study of the leading multi-trait GWAS methods, providing a web application and open-source software program implementing our simulation framework for further benchmarking of multi-trait GWAS methods.

Motivated by our comparison results, we develop novel methodology and present a series of multi-trait analyses. We perform multi-trait genome-wide analyses on publicly available GWAS summary statistics on 19 traits – metabolic, anthropometric and psychiatric. We develop and apply two summary statistic methods: one that has increased power to detect pleiotropic effects on multiple traits, and one that is more powerful for detecting heterogeneous genetic effects.

Polygenic risk scores (PRS) are now a commonly used tool for performing phenotype prediction from genetics, assessing the genetic aetiology underlying diseases, and testing for shared genetic aetiology among traits. Using UK Biobank data, we explore

the predictive ability of PRS computed across multiple traits for Major Depressive Disorder (MDD). The MDD PRS itself has so far offered modest prediction of MDD case/control status; we explore the use of PRS built on traits correlated with MDD to improve predictive ability. We build main effect and interaction models, using both AIC and BIC stepwise variable selection, and cross-validation, to establish the most predictive models.

Acknowledgements

I would like to express my appreciation and thanks to my supervisors, Dr Paul O'Reilly and Professor Cathryn Lewis, for their invaluable contributions and guidance throughout the last three years. Extended thanks go to the Statistical Genetics Unit, and colleagues at the SGDP Centre, for stimulating scientific discussions that encouraged my development as a researcher. Finally, I would like to thank my family for their unwavering support, and my partner, Matt, for his encouragement and much-needed wit.

I would like to thank the Medical Research Council for funding my research.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Overview

Genome-wide association studies (GWAS) traditionally adopt a univariate approach, focusing on a single phenotype of interest (Gieger et al., 2011; Teslovich et al., 2010; Willer et al., 2008; Morris et al., 2012). In recent years numerous methods have been proposed that model multiple phenotypes simultaneously to investigate their joint association with single-nucleotide polymorphisms (SNPs) (O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Ferreira and Purcell, 2009; Klei et al., 2008; Aschard et al., 2014; Marchini et al., 2007; Stephens, 2013; Nath and Pavur, 1985; Zhou and Stephens, 2014; Korte et al., 2012; Segura et al., 2012; Schifano et al., 2013) and to increase the statistical power for discovery of susceptibility loci. There are also methods that, in addition, model multiple SNPs jointly (Bottolo et al., 2013; Servin and Stephens, 2007; Casale et al., 2015) which can lead to increased power to detect susceptibility loci due to reduced residual variation.

Methods that model multiple phenotypes jointly could be utilised to uncover relationships between phenotypes that do not have established relationships, or could be used to modify existing phenotype definitions. The modelling of multiple SNPs jointly also allows the exploration of the biological relationships underpinning correlated phenotypes. This could highlight genetic pathways important to certain groups of phenotypes, and motivate investigation into their shared biology.

So far there has not been an extensive, systematic comparison of multi-trait GWAS methods and thus there is no clear overall preference or guidance on when different multivariate methods should be implemented in order to optimise discovery.

Moreover, many of the methods have not been tested under certain scenarios and so there is a need for a comprehensive simulation study comparing the relative performance of these methods, which may lead to improved methodology. Largely due to lack of clarity as to the relative performance of multi-trait GWAS methods compared to the univariate approach, there have been few large-scale multi-trait GWAS performed (Adhikari et al., 2015, 2016; Kauwe et al., 2014). Producing and applying the most powerful method to GWAS will likely lead to new discoveries and novel drug targets.

Polygenic risk scores (PRS) (Purcell et al., 2009; Euesden et al., 2015; Dudbridge, 2013) have so far only been applied to pairs of phenotypes. Often a PRS is built in one phenotype then used to predict either itself, to assess the predictive ability of the PRS, or to predict another related phenotype to assess the shared genetic aetiology between traits. By using PRS built on multiple phenotype predictors that correlate with the trait we wish to predict, new insights into the shared biology between phenotypes can be gained, as well as informing phenotype stratification.

1.2 Genome-wide association studies

Genome-wide association studies (GWAS) have become a successful tool for identifying associations between common genetic variants and many different diseases and traits. GWAS test hundreds of thousands of genetic variants, typically single nucleotide polymorphisms (SNPs), across the genome for their association with a phenotype of interest. Since the emergence of GWAS in 2007 (Burton et al., 2007), thousands of SNPs have been found to be associated with a range of complex traits, under the common disease common variant (CDCV) hypothesis

(Reich and Lander, 2001) which states that many frequent, low risk variants contribute to diseases with high population prevalence.

The formation of large consortia, such as the PGC (Psychiatric Genomics Consortium) and GIANT (Genetic Investigation of ANthropometric Traits) for the meta-analysis of multiple data sets has led to increased sample sizes and statistical power, and thus even greater discovery of genetic variants using the GWAS approach. Psychiatric traits highlight the successes and failures of GWAS. More than 100 genetic variants have been found to be associated with schizophrenia (**Figure 1;** Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014); in contrast, for Major Depressive Disorder (MDD) there are few known genetic determinants to date (Ripke et al., 2013; Converge Consortium, 2015; Hyde et al., 2016). This difference in findings could be due to several factors, such as differences in the quality of phenotyping, in the statistical power of the studies or due to incorrect disease model assumptions. A recently published study using data from the genetic testing company 23andMe (Hyde et al., 2016), and meta-analysed with the PGC MDD GWAS (Ripke et al., 2013), on $326{,}113$ individuals ($84{,}847$ cases) has led to the discovery of 4 independent loci associated with MDD. This highlights the need for large sample sizes when studying heterogeneous traits in order to provide sufficient statistical power to detect causal genetic associations.

Methodology development for increasing the statistical power of genetic association studies is extremely important for heterogeneous traits and those with small sample size GWAS, in order to maximise the discovery potential and thus provide insights into the biology underlying these diseases.

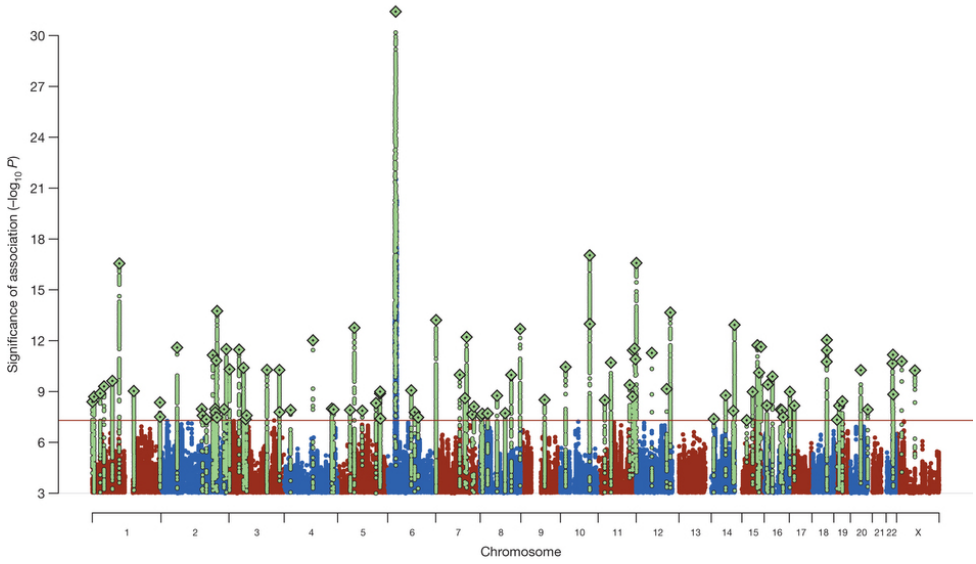

Figure 1. Manhattan plot for the latest schizophrenia (SCZ) GWAS from the Psychiatric Genomics Consortium (PGC), illustrating the 108 known loci (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

1.2.1 Multi-trait GWAS

GWAS are most commonly performed on a single phenotype of interest, or across many related traits, such as lipids (Teslovich et al., 2010; Willer et al., 2008), but with each trait analysed separately. The association between a SNP and a phenotype is typically tested by performing linear regression for continuous traits and logistic regression for case/control phenotypes, where each SNP is regressed on the phenotype separately. The association is considered significant if the *P*-value is below the genome-wide significance threshold of 5×10^{-8}. Although this has been the primary method of analysis for most GWAS, it allows only one phenotype to be considered at a time, despite the likelihood that SNPs often associate with many traits. In recent years, numerous multivariate methods have been proposed; these model multiple phenotypes simultaneously to investigate their joint association with a SNP. These methods allow shared genetic architecture and correlation between phenotypes to be exploited, potentially increasing the power to detect true genotype-

phenotype associations. There have been some applications of multi-trait GWAS methodology in the literature (Kauwe et al., 2014; Adhikari et al., 2015, 2016), however they have so far only been applied to relatively small sample sizes, thus making it difficult to ascertain the performance gains from current applications. However, large multi-trait resources, such as the UK Biobank and planned US Biobank, should help to facilitate such analyses.

There are two types of GWAS method that test multiple phenotypes jointly: single SNP methods (O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Bolormaa et al., 2014; Ferreira and Purcell, 2009; Klei et al., 2008; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007; Nath and Pavur, 1985; Zhou and Stephens, 2014; Korte et al., 2012; Segura et al., 2012; Schifano et al., 2013; Casale et al., 2015), which test the joint association of a set of phenotypes SNP-by-SNP, and multi-SNP methods (Bottolo et al., 2013; Servin and Stephens, 2007; Zhou and Stephens, 2012; Kim et al., 2016; Casale et al., 2015), which test the association between multiple phenotypes and multiple SNPs simultaneously. Within the single SNP category, there are two further types: those that use summary statistics from existing univariate analyses (O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Bolormaa et al., 2014) and those that use individual-level genotype-phenotype data (O'Reilly et al., 2012; Ferreira and Purcell, 2009; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007; Nath and Pavur, 1985; Klei et al., 2008; Zhou and Stephens, 2014; Korte et al., 2012; Segura et al., 2012; Schifano et al., 2013; Casale et al., 2015).

1.2.2 Summary statistic methods

By applying multi-trait GWAS methods based on summary statistics, previous studies on separate phenotypes can be repurposed to conduct a multi-trait study without

additional data collection. Furthermore, often much larger sample sizes of summary statistics are available than individual-level data. Currently, the most common approach to multi-trait GWAS is to perform separate univariate GWAS for each phenotype, and adjust the *P*-values to account for multiple testing (The International Consortium for Blood Pressure, 2011), though often this correction is not applied and the usual 5×10^{-8} genome-wide significance threshold is applied. A Bonferroni correction (Bland and Altman, 1995) can be applied to the *P*-value threshold α, to obtain a corrected significance threshold of $\alpha' = {}^{\alpha}/_{N}$, for N tests. However, this is a highly conservative correction as this assumes that all the univariate analyses were independent, and thus does not account for the correlation between traits. A simple and less conservative way of correcting for multiple testing is to apply a standard Šidák (Sidak, 1968) correction to the minimum of the univariate *P*-values, incorporating the effective number of independent tests, which is determined by the trait correlations. A formula for the effective number of tests is given by Nyholt (Nyholt, 2004), based on the eigen-decomposition of the phenotype correlation matrix. The min-P method (O'Reilly et al., 2012) uses Nyholt's effective number of tests calculation, based on the number of GWAS results being tested and the correlation of their results, and then performs a Šidák correction on the minimum *P*-value from a set of univariate GWAS results at a SNP. This method was used by the authors to benchmark their individual-level based method, MultiPhen (O'Reilly et al. 2012).

TATES (Trait-based Association Test that uses Extended Simes procedure) (van der Sluis et al., 2013) is a method that has been developed exclusively for utilising existing summary statistics from univariate analyses. The method requires the phenotype correlation matrix and the *P*-values obtained from univariate analyses on each of these phenotypes; so for K phenotypes, the method requires the $K \times$

K phenotype correlation matrix and the K univariate *P*-values for each SNP. Rather than use the phenotype correlation matrix to deduce the number of independent tests, TATES transforms this trait correlation matrix into a corresponding *P*-value correlation matrix via a sixth order polynomial that was determined by simulation, and uses the eigen-decomposition of this *P*-value correlation matrix. The univariate *P*-values are weighted according to this eigen-decomposition, and the minimum of these weighted *P*-values is chosen as the corrected *P*-value for the joint association. TATES has been shown to outperform univariate analyses and other multi-trait methods that use individual-level data, such as MultiPhen (O'Reilly et al., 2012), in certain scenarios (van der Sluis et al., 2013). It is unclear whether this higher power is generalisable, or a characteristic of the modelling scenarios imposed in the simulations. Summary statistic methods are more easily applied, since complete genotype-phenotype data across multiple traits are often not available, and there are now growing resources of publicly available summary data. However, by using only the *P*-values from univariate studies, they are not able to fully exploit the correlation structure between the phenotypes.

Multi-trait methods have also been developed that use the signed t-values in a meta-analysis across traits. These methods are able to exploit the correlation between phenotypes via the correlation between the t-values. The S_{Het} and S_{Hom} methods (Zhu et al., 2015) combine the t-values from multiple traits, and across multiple cohorts, in a sample size and t-value correlation weighted meta-analysis. The S_{Hom} method performs a meta-analysis across all traits under study; the S_{Het} method, however, meta-analyses subsets of traits, with these subsets being determined by their univariate t-values and some user specified threshold. For each SNP, the association between the SNP and all traits with an absolute univariate association t-value $> T$ for some threshold T is tested. This analysis is then repeated across a range of different thresholds, or is performed for one threshold if specified by the user. The default is to

order the traits by their univariate t-values, recursively performing associations between subsets of the traits and the SNP, starting with the trait with the largest absolute univariate t-value, and progressively adding each trait until all traits are included in the association. It has been shown that the S_{Hom} method has greater power to detect homogenous genetic effects (or pleiotropic genetic effects), but when heterogeneous genetic effects exist the S_{Het} method is preferred as subsets of traits are tested for their association with the genetic variant (Zhu et al., 2015). Another standard approach to meta-analysis across traits has been proposed (Bolormaa et al., 2014), which weights the t-values only by their t-value correlations, whereas S_{Het} and S_{Hom} perform t-value correlation and sample size weighted meta-analyses. This method has been identified as performing best under homogenous genetic effects due to having a large number of degrees of freedom compared to the other two tests (Zhu et al., 2015).

The public release of GWAS summary statistics across multiple traits motivates the development of multi-trait GWAS methods for increasing discovery and identifying pleiotropic loci. A Bayesian approach to performing multi-trait GWAS on summary statistic data has been recently proposed (Zhu and Stephens, 2016), where a summary statistic specific likelihood is developed and implemented for the detection of genetic variants. Furthermore, the recently published method GWIS (genome-wide inferred statistics) (Nieuwboer et al., 2016), can construct summary statistic data for compound traits, such as BMI, from the individual component summary statistics, expanding the utility of summary data to phenotypes for which there was no previously available resource. The utility of summary statistics has been further explored in other genetic analysis tools, such as polygenic risk scores (PRS) for quantifying genetic risk of disease (Purcell et al., 2009; Euesden et al., 2015; Dudbridge, 2013), LD Score regression for calculating the genetic correlation between traits (B. Bulik-Sullivan et al., 2015) and for estimating heritability (B. K.

Bulik-Sullivan et al., 2015). These tools use summary data in order to gain further insight into the genetic aetiology of traits, without the requirement for large resources of individual-level data. A review of current summary statistic methodology has recently been performed (Pasaniuc and Price, 2016). The authors conclude that, although there are certain limitations to using summary data, for example loss of accuracy for applications such as imputation, given the large sample sizes of available summary data, methods exploiting summary statistics are often preferable to their individual-level counterparts. An online resource of GWAS summary statistics has recently been developed (Staley et al., 2016), allowing the vast amount of summary statistic data to be easily extracted across traits, which will further facilitate summary statistic based analyses.

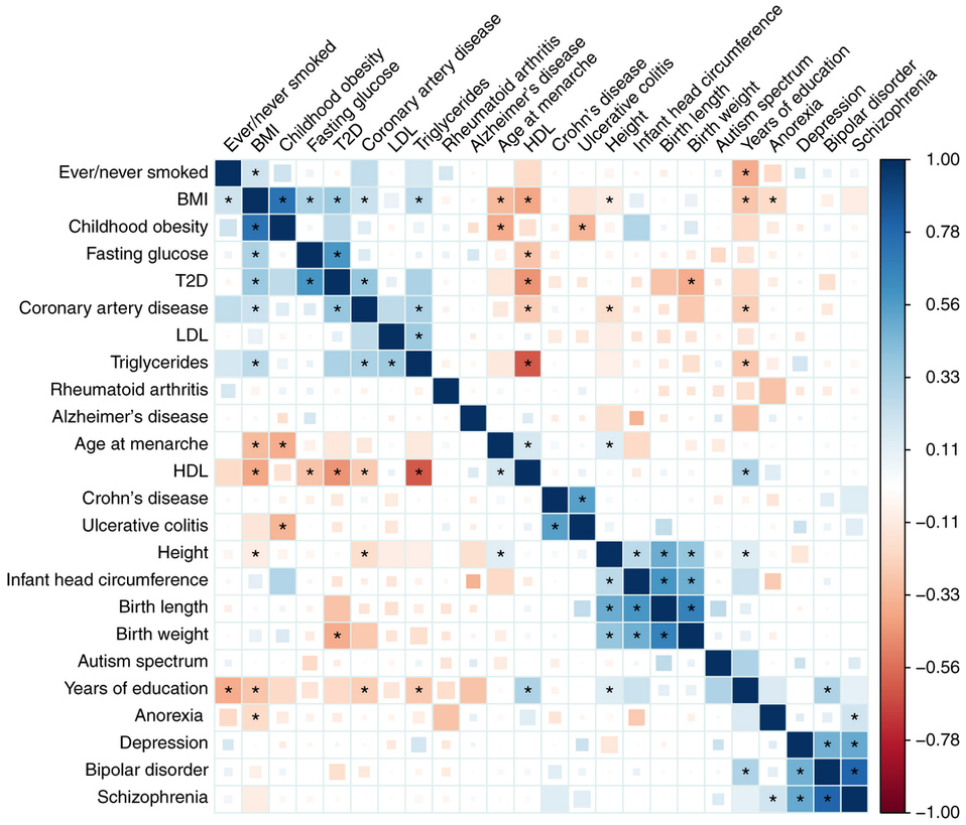

Figure 2. Genetic correlation 'atlas' plot from the application of the LD score regression method (B. K. Bulik-Sullivan et al., 2015) to publicly available summary statistic data (B. Bulik-Sullivan et al., 2015).

1.2.3 Individual-level methods

Methods that utilise individual-level genotype-phenotype data have been the most commonly developed multi-trait method in recent years. Ferreira and Purcell (Ferreira and Purcell, 2009) proposed the use of canonical correlation analysis (CCA) for modelling multiple phenotypes jointly. The statistical method CCA tests the association between two sets of variables, thus could be applied to multiple SNPs, but here it is presented as a single-SNP, multi-trait method. This method is equivalent in power to MANOVA (Multivariate ANalysis Of VAriance) (Nath and Pavur, 1985; O'Reilly et al., 2012). Another method that has been shown to be approximately equivalent to CCA and MANOVA for common genetic variants is MultiPhen (O'Reilly et al., 2012). MultiPhen reverses the regression of usual univariate analyses, treating the SNP as outcome and multiple phenotypes as predictors, in an ordinal logistic regression model. What makes CCA and MultiPhen differ is the assumed distribution of the outcome; CCA assumes normality of residuals, whereas MultiPhen models the genotype as an ordinal random variable. Both methods find the linear combination of phenotypes most associated with the SNP, and a likelihood ratio test is performed on the full model versus that with no phenotype predictors. These methods mostly outperform the univariate approach, particularly when the direction of genetic effect is not concordant with the phenotypic correlation, however when the direction is the same, their power appears greatly reduced compared to the univariate approach (O'Reilly et al., 2012). Aschard *et al.* (Aschard et al., 2014) developed a method that performs a principal components analysis (PCA) on the individual-level phenotype data. For K phenotypes there are K principal components, which are tested for association with a SNP. For each principal component (PC), a univariate analysis is performed with SNP as predictor and PC as outcome, resulting in K univariate models for each SNP. Since PCs are

independent, the chi-squared statistics from each univariate model can be summed, resulting in a single *P*-value representing the joint association between the SNP and all the PCs. The combined-PC method was shown to have similar power to MultiPhen across the scenarios considered, but MultiPhen had type I error inflation when analysing 50 phenotypes (Aschard et al., 2014). The authors suggest that it is not always best to consider only the top few PCs, and that the effect of the SNP on the phenotype can be spread across all or a subset of the PCs. Disentangling which PCs are detecting the genetic effect could lead to a method with more power than the proposed combined-PC approach due to a reduction in the degrees of freedom of the test.

Bayesian approaches to GWAS have also been proposed for single-SNP, multi-trait analyses (Stephens, 2013; Marchini et al., 2007), but have had limited application in the field so far due to their computational intensity. The package BIMBAM (Bayesian IMputation-Based Association Mapping) (Stephens, 2013; Servin and Stephens, 2007), performs Bayesian multivariate regression to test for association, partitioning the phenotypes according to the effect of the genetic variant, and using Bayes factors to assess the association between the groups of phenotypes and a SNP. Different association scenarios are modelled, such as separate direct effects between the SNP and all traits or indirect effects between the SNP and a subset of the traits, captured in a Directed Acyclic Graph (DAG) framework, and the support for each model relative to the null hypothesis is computed. The BIMBAM package also performs multi-SNP analysis. The software package SNPTEST (Marchini et al., 2007), for performing SNP association testing in GWAS, now incorporates Bayesian association testing for multiple phenotypes, accounting for genotype uncertainty. The model implemented is a Bayesian multivariate linear model with a conjugate prior, consisting of an inverse Wishart prior on the error covariance matrix, and a matrix normal prior on the genetic effect parameters (A. P. Dawid, 1981).

Mixed models have also been considered as a way of testing for joint association. Zhou and Stephens (Zhou and Stephens, 2014) proposed the use of multivariate linear mixed models (mvLMMs), incorporated into the software GEMMA (Genome-wide Efficient Mixed Model Association) (Zhou et al., 2013; Zhou and Stephens, 2012), for testing the association between a single SNP and many correlated phenotypes, while controlling for population stratification. mvLMMs have previously been used to estimate heritability (Zhou and Stephens, 2014; Price et al., 2011), as well as studying pleiotropy and genetic correlation (Lee et al., 2012; Zhou and Stephens, 2014). mvLMMs are also a useful tool for GWAS as they can adjust for population stratification and cryptic relatedness between individuals (Zhou and Stephens, 2014). Algorithms for fitting mvLMMs can be highly computationally intensive, such as in GCTA (Genome-wide Complex Trait Analysis) (Yang et al., 2011) and therefore they are not practical for conducting likelihood ratio tests (LRTs) in a GWAS setting (Zhou and Stephens, 2014). GEMMA, however, is able to fit mvLMMs and perform LRTs in GWAS by extending existing univariate linear mixed model (LMM) methods, (Zhou and Stephens, 2012; Lippert et al., 2011; Pirinen et al., 2013), and applying them in a multivariate context. Although GEMMA provides a more computationally efficient alternative, it is still only practical genome-wide for no more than 10 phenotypes, and sample sizes up to $50{,}000$ (Zhou and Stephens, 2014). Segura *et al.* (Segura et al., 2012) proposed a mixed-model approach that performs multi-locus association testing, for use in structured populations. It uses stepwise mixed-model regression, and the authors explore two model selection criteria: an extended Bayesian information criterion (EBIC) and a multiple Bonferroni criterion (mBonf). It was shown that this method performed better than alternatives for structured samples and when studying phenotypes for which several SNPs have moderate to large effect sizes (Segura et al., 2012). Schifano *et al.* (Schifano et al., 2013) proposed a scaled marginal model (SMM) for both testing and estimating the effect of a SNP on many phenotypes in case/control studies. The authors found that

their method, SMAT (Scaled Multiple-phenotype Association Test), has higher power than univariate analyses, as well as higher power than other multi-trait methods when the phenotypes are positively correlated and when they capture an underlying trait in the same direction (Schifano et al., 2013).

1.2.4 Multi-SNP methods

Many methods have been proposed in the literature for testing the association of multiple phenotypes accounting for the trait correlation structure, but in most cases correlated sets of traits are tested for association SNP-by-SNP. Multi-SNP methods aim to increase power by reducing the residual variance by including other genetic variants as predictors in the model. The correlation structure between genetic variants across the genome is highly spatial, in that SNPs physically close are often highly correlated, whereas distant SNPs tend to be uncorrelated due to genetic linkage (The International HapMap 3 Consortium, 2010). The correlation structure between genetic variants is measured by linkage disequilibrium (LD) (Reich et al., 2001). By modelling multiple SNPs jointly, and controlling for this correlation structure, multi-SNP methods may have increased power to identify susceptibility loci. Most multi-SNP methods adopt a Bayesian approach; in a review of Bayesian approaches to genetic association testing (Stephens and Balding, 2009), Stephens and Balding discuss the use of Bayes factors over *P*-values, as well as the choice of prior distribution. GUESS (Graphical Unit Evolutionary Stochastic Search Algorithm) (Bottolo et al., 2013) is one such multi-SNP Bayesian method. This method is an amalgamation of Bayesian variable selection (BVS) (Guan and Stephens, 2011) procedures, to extend single-SNP GWAS analyses, and multi-trait modelling methods. GUESS uses linear regression to test for association, with SNP predictors and phenotype outcomes, along with an evolutionary stochastic search algorithm (Bottolo and Richardson, 2010) that searches genome-wide for subsets of SNPs

associated with multiple phenotypes in order to reduce dimensionality. GUESS was shown to outperform the Bayesian multi-trait association testing of SNPTEST in both single-phenotype and multi-phenotype simulations (Bottolo et al., 2013), but GUESS is highly computationally intensive, partly due to the stochastic search algorithm component of the method. The multi-SNP methods can also be utilised for single-SNP analyses. CCA, MultiPhen, SNPTEST and BIMBAM have similar power for the modelling scenarios considered in a recent comparison study (Galesloot et al., 2014).

A common misconception in the field of multi-trait GWAS is that there must exist pleiotropic effects between the genetic variant and most or all of the phenotypes under study for a gain in power to be achieved by the multivariate approach. However, gains in power can be achieved even if only one of the traits exhibits a causal relationship with the genetic variant (O'Reilly et al., 2012; Zhou and Stephens, 2014). A multivariate model assesses the overall association between the SNP and the phenotypes in the model, as compared to a model with none of the phenotypes included, and thus conclusions about the effects on the individual phenotypes cannot be made directly. Including additional phenotypes that correlate with the associated phenotype can reduce the residual variance because the correlated phenotypes likely share a high proportion of risk factor effects. Thus, the association between the variant and the one associated phenotype explains a high fraction of remaining residual variance, resulting in high statistical power.

Zhou and Stephens (Zhou and Stephens, 2014) state that there is no one most powerful method, reiterated in a recent comparison study (Galesloot et al., 2014), but that instead the different methods should be viewed as complementary. This highlights the importance of gaining a deeper understanding into the performance of

these methods to identify the scenarios in which each method should be adopted in order to maximise discovery potential.

1.3 Assessing genetic aetiology

It has been established that most common diseases and traits are polygenic whereby hundreds or even thousands of genetic variants have small contributing effects across the genome (Wood et al., 2014; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Global Lipids Genetics Consortium, 2013). Larger sample sizes are sought when performing GWAS on such traits in order to detect the small genetic effects that associate with the disease/trait outcome. This is particularly pertinent for traits such as Major Depressive Disorder (MDD), in which high heterogeneity in the phenotype contributes further to produce low power of discovery. Methods have been developed to exploit the polygenicity of complex traits to optimise phenotype prediction from genetics and to gain understanding of their genetic architecture. Polygenic risk scores (PRS) are one such tool that aggregate information on genetic risk across the genome (Purcell et al., 2009; Euesden et al., 2015; Dudbridge, 2013). In addition, the LD Score regression method (B. Bulik-Sullivan et al., 2015) allows the genetic correlation between traits to be determined, providing information about the genetic overlap between traits, and methods such as GCTA (Yang et al., 2011) and LD Score regression (B. K. Bulik-Sullivan et al., 2015) allow the heritability and co-heritability of traits to be estimated. Furthermore, the identification of pleiotropic SNPs (SNPs that effect more than one phenotype) is likely to become an increasingly important area of research as we seek to further our understanding of complex trait genetics.

1.3.1 Pleiotropy

The term pleiotropy historically refers to a single gene influencing multiple, very different, traits (Solovieff et al., 2013). In the context of complex disorders, which are influenced by many small genetic effects across the genome (Visscher et al., 2012), we can also consider pleiotropy at the SNP level, and so the term has evolved to also consider effects on multiple, related traits. Therefore, in the context of investigating complex disorders, for which SNP effects are unlikely to be unique to particular traits, SNPs are considered pleiotropic even when influencing similar traits, such as LDL and triglycerides. Identifying pleiotropic SNPs can lead to greater understanding of the underlying biological network between traits, and identify biological pathways enriched for effects on clusters of traits for further investigation.

A recent review paper on the topic of pleiotropy (Solovieff et al., 2013) considered how different types of pleiotropy can arise, and noted the importance of distinguishing between them in order to gain biologically meaningful conclusions as to the shared biological mechanisms between traits. Solovieff *et al*. highlight what they classify as 'cross-phenotype' associations, where a SNP expresses an association with more than one phenotype, and the important distinction with pleiotropic SNPs. For a SNP to be pleiotropic, they state, it must have a genuine effect on more than one trait, rather than merely having an observed association with multiple traits. Previous studies have supported the existence of pleiotropic SNPs through the identification of loci that affect more than one trait, such as the *TYK2* locus in Crohn's disease and psoriasis (Franke et al., 2010; Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, 2010 in Solovieff et al., 2013). In addition, polygenic risk score and LD score regression analyses have established shared genetic aetiology between a huge range of traits (Power et al., 2015; Bulik-Sullivan et al., 2015; Krapohl et al., 2015), for example

between schizophrenia and bipolar disorder (Purcell et al., 2009), as well as type 2 diabetes and hypertension (Lee et al., 2012 in Solovieff et al., 2013).

Pleiotropy can arise in many forms, and distinguishing between them is important for understanding the biological implications. **Figure 3** shows a graphical representation of the different forms of pleiotropy (Solovieff et al., 2013).

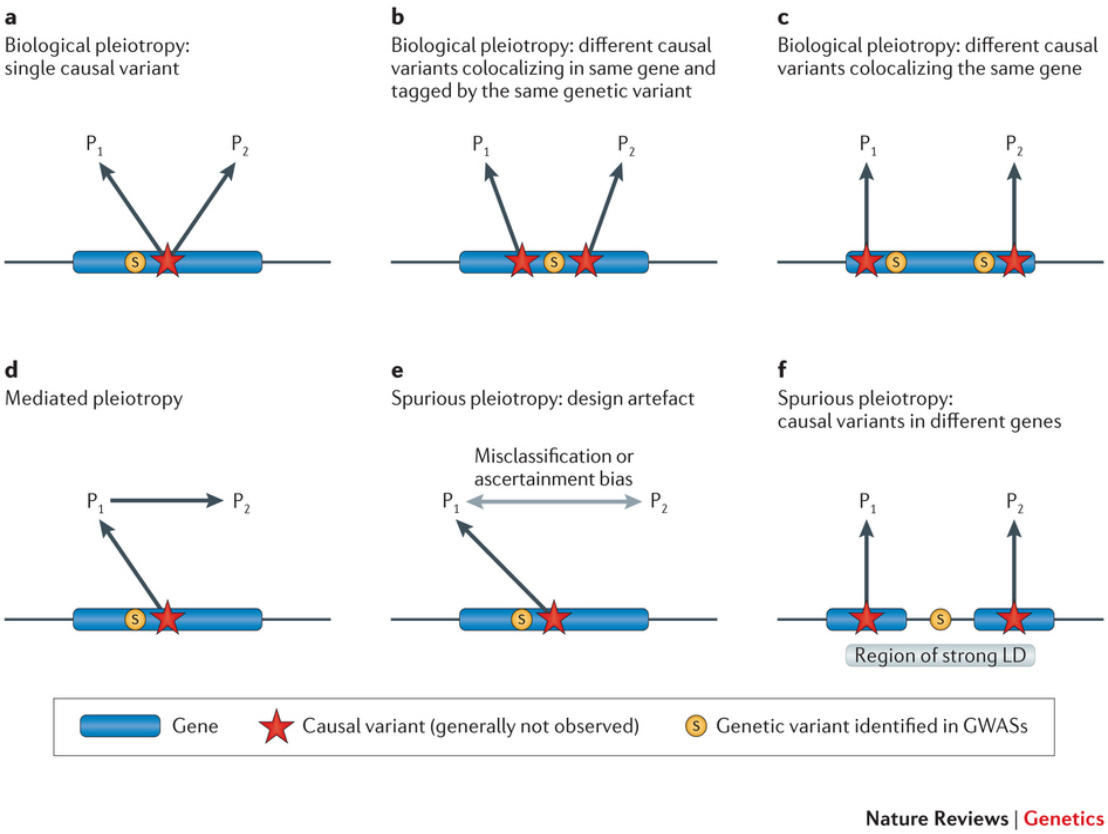

Figure 3. A visual representation of the different forms of pleiotropy as illustrated in a recent review paper (Solovieff et al., 2013).

Pleiotropy can refer to a single causal variant with effects on both traits, P_1 and P_2 (**Figure 3a**), but can also refer to two causal variants that are in the same gene or locus that are tagged by the same genetic variant (**Figure 3b**). However, two causal variants in the same gene or locus that are tagged by two separate genetic variants can also be classified as pleiotropic SNPs (**Figure 3c**). These three types of pleiotropy are classified as biological pleiotropy, where the causal variant(s) are

having direct effects on the phenotypes of interest. The other type of pleiotropy is mediated pleiotropy, whereby a causal variant, tagged by a genetic variant, has an effect on one phenotype P_1, which then has a downstream effect on another phenotype P_2; we say that the genetic effect on P_2 is mediated by P_1 (**Figure 3d**). While biological and mediated pleiotropy have different biological implications, they can be difficult to differentiate using GWAS results alone. The final two representations of pleiotropy in **Figure 3** refer to spurious pleiotropy, where there appears to be a pleiotropic effect when one does not exist. Spurious pleiotropy can arise through misclassification of phenotypes or ascertainment bias (**Figure 3e**), or, when in a region of high linkage disequilibrium, there is a single genetic variant tagging two causal variants in different genes (**Figure 3f**).

The main challenge in the analysis of pleiotropic genetic variants is in distinguishing between the different types, and translating this knowledge into meaningful biological findings. Mendelian randomisation is a technique that aims to identify whether there exists direct effects from the genetic variant to each phenotype, or whether there exists an intermediate effect on one phenotype that has a downstream effect on another (Davey Smith and Hemani, 2014). This method therefore attempts to distinguish between biological and mediated pleiotropy, as illustrated in **Figure 3** above (Solovieff et al., 2013). Another approach to further characterising pleiotropy is the PRIMe pleiotropy index (Huang et al., 2011). This method is able to first detect regions where there exist multiple associations from published GWAS, and then define a pleiotropy index based on the number of traits with which the region is associated. While this approach does not make any distinction between pleiotropy types, it provides a way of quantifying the degree of pleiotropy and highlighting pleiotropic 'hotspots' for further follow up analysis. In the context of multi-trait GWAS, while most methods are not optimised for the detection of pleiotropic variants, nor do they require a pleiotropic effect to gain power over the univariate approach, they do

have the potential to further describe the effect of a genetic variant on multiple traits. However, most of those described in this thesis (with the notable exception of mv-BIMBAM (Stephens, 2013)) are focused on increasing statistical power for detecting variants affecting some number of the traits, and not on establishing the type effect or form of pleiotropy. The study of pleiotropic variants, while challenging, is likely to be instrumental in both increasing our knowledge of related diseases and traits, and in the development of effective treatments.

1.3.2 Polygenic risk scores

A polygenic risk score (PRS) is an effect-size weighted sum of risk alleles present across the genome of an individual, and thus quantifies the genetic load of disease for an individual (Purcell et al., 2009; Euesden et al., 2015; Dudbridge, 2013). PRS are calculated according to a number of *P*-value thresholds, which determine the number of SNPs to be included in the score based on the univariate GWAS *P*-values for each SNP. The most predictive score of the studied trait among those calculated at different thresholds is typically chosen for subsequent analyses. PRS provide a useful tool for assessing how at risk an individual is for a particular disease, as well as for determining the polygenicity of traits. They were originally presented in the psychiatric genetics field (Purcell et al., 2009), but have since been applied across numerous human traits (Krapohl et al., 2015; Selzam et al., 2016; Power et al., 2015; Hung et al., 2015; Vassos et al., 2016). They are particularly useful as a way of leveraging the results of an underpowered GWAS. In the first study to investigate the concept of polygenic risk, the authors found that the PRS for schizophrenia (SCZ) both significantly predicts SCZ case/control status and also case/control status of Bipolar Disorder (BPD) (Purcell et al., 2009). Since then, PRS have been commonly applied to assess shared genetic aetiology across traits, and phenome-wide PRS

approaches have been adopted by performing such analyses systematically across many traits (Krapohl et al., 2015).

PRS method development has also been an active area of research interest in the field; the software PRSice calculates high-resolution PRS at hundreds of *P*-value thresholds in order to find the most predictive threshold, while the AVENGEME software exploits theory developed in Dudbridge 2013 (Dudbridge, 2013) to estimate heritability, co-heritability and the proportion of causal variants from polygenic risk scores (Palla and Dudbridge, 2015).

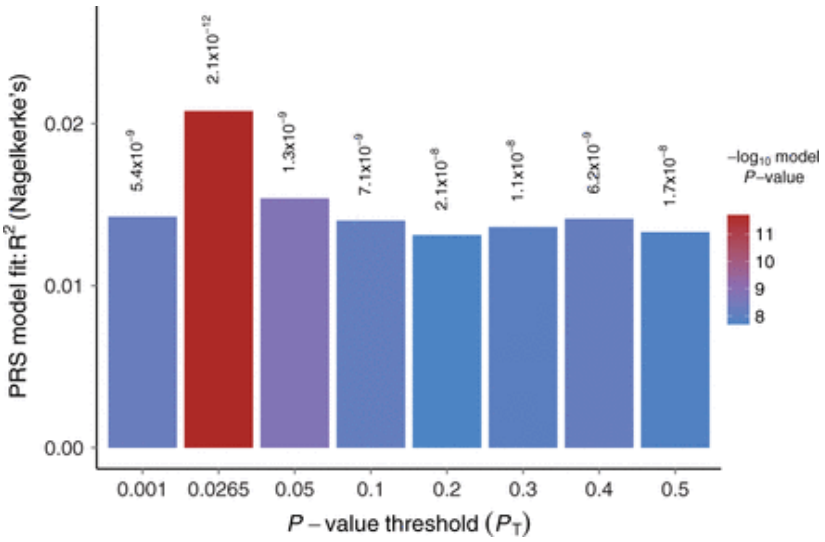

Figure 4. Bar plot from the high-resolution polygenic risk score (PRS) method PRSice; the PRS for SCZ at the $P_T = 0.0265$ threshold is most predictive of SCZ case/control status (Euesden et al., 2015).

Despite PRS now being routinely applied across traits to assess shared genetic aetiology, this has been performed on a pairwise basis where a PRS built on one trait is used to predict another trait. A multivariate approach, whereby PRS for multiple traits are used to predict disease outcomes, has yet to be explored. The genetics of one trait do not act independently of the genetics of another, and given that the biological network underlying multiple traits is a highly connected network, interaction

between genetic markers occurs. Genetic interaction is most commonly tested statistically between a small number of SNPs or genes, rather than across the whole genome, due to the computational complexity of modelling such a large number of interactions. The interaction between PRS could thus provide broad insight into the genetic interaction influencing common traits.

1.3.3 Major Depressive Disorder

Major Depressive Disorder (MDD) is a debilitating psychiatric disorder estimated to affect more than 350 million people worldwide ('WHO | Depression', Fact Sheet; Whiteford et al., 2013). Given the high heritability ($h^2 = 37\%$) (Sullivan et al., 2000), there have been intense research efforts to establish the genetic underpinnings of this common, complex disorder.

Most GWAS performed on MDD to date have had limited success. The first genetic loci found to be associated with MDD was the result of whole-genome sequencing of around $5{,}000$ Chinese women with recurrent MDD (Converge Consortium, 2015). This population was chosen due to under-diagnosis of MDD in China, thus, in addition to restricting to only recurrent cases of MDD, provided a more homogenous phenotype. There has also been a study into the trade-off between sample size and read depth in whole-genome sequencing; it was found that larger sample sizes at lower read depth provide greater statistical power (Pasaniuc et al., 2012), which was the approach adopted in this study. More recently, the largest study of MDD to date (Hyde et al., 2016) identified 4 independent loci from a GWAS on $326{,}113$ individuals ($84{,}847$ cases). The study was performed using data from the consumer genetic testing company 23andMe, and was meta-analysed with the PGC MDD GWAS (Ripke et al., 2013). The MDD phenotype obtained from 23andMe is likely to be more

heterogeneous than provided by a clinical sample, due to MDD diagnoses being determined by a simple online question with no confirmation of diagnosis; this heterogeneity, however, is likely to be at least somewhat negated by the large sample size.

Due to the limited success of MDD GWAS to date, PRS have been frequently applied, though the predictive ability of the MDD PRS is limited (Ripke et al., 2013). The heterogeneity of MDD suggests that there exist many MDD subtypes, which would explain why there has been limited success to date using the dichotomous MDD definition, as defined by the DSM-IV criteria. PRS have been used to assess the shared genetic aetiology between MDD and associated outcomes such as childhood trauma and stressful life events (Mullins et al., 2016). In addition, this approach has been adopted in an attempt to identify subtypes of MDD by analysing symptomatic data (Milaneschi et al., 2016; Levine et al., 2014; Okbay et al., 2016). Multiple PRS on sub-types or on correlated traits have not, however, been used to simultaneously predict MDD case-control status. The multi-dimensionality of MDD suggests that there are multiple contributing factors (both genetic and environmental), and that interaction between these factors across the genome could associate with disease status. **Chapter 5** investigates prediction models of MDD based on PRS computed from multiple MDD relevant traits.

1.3.4 Phenotype stratification

Many complex disorders are heterogeneous, meaning that within individuals defined as cases, there is variability in the exhibited phenotype(s). A recent paper, for example, examined the polygenicity of MDD and two known clinical subtypes: typical and atypical, in order to gain a greater understanding into the MDD phenotype (Milaneschi et al., 2016). By grouping individuals only by their case/control status, we

potentially introduce noise into the data, and over-simplify the disorder, reducing information, by considering it as a binary trait. Phenotype stratification, where a phenotype considered to be one disorder is dissected into its more elementary components, is a particularly important objective of personalised medicine. In MDD, for example, some individuals experience no success with many different types of anti-depressant. One contributing factor could be that there are many underlying forms of MDD, and that currently available anti-depressants are not targeting these subtypes specifically. If we are able to stratify complex disorders into more fundamental components, this may lead to novel drug targets specific to these subtypes and greatly improve the prognosis of complex traits. Multi-trait analyses could hold the key to stratifying complex disorders, by assessing the shared genetic aetiology between correlated traits, and by performing multi-trait analyses on endo-phenotypes to elucidate the observed phenotypic variation by identifying distinct sub-types.

1.4 Chapter Outline

In **Chapter 2** we develop and present a multivariate simulation framework for modelling the effect of a SNP on multiple correlated traits in the context of multi-trait GWAS. We construct a large range of modelling scenarios that aim to capture different aspects of genotype-phenotype associations, by considering the extra modelling challenges of multi-trait GWAS. The simulation framework is provided as a command-line software package with associated R shiny web application for simulation of multivariate genetic datasets, and for the testing and development of multi-trait GWAS methodology.

Chapter 3 utilises the simulation framework built in **Chapter 2** to perform a comprehensive comparison of the leading multi-trait GWAS methods. We compare methods that utilise univariate GWAS summary statistics, as well as those methods that exploit individual-level genotype-phenotype data directly. Our findings expose the similarities and differences across different approaches to multi-trait GWAS and act as a guide to other researchers involved in the genetic analysis of multiple traits. Furthermore, our findings highlight areas for potential improvement in the development of novel methodology.

Motivated by our findings of **Chapter 3**, in **Chapter 4** we perform multi-trait GWAS on a variety of traits (both quantitative and binary) using publicly available GWAS summary statistics. We develop and apply two summary statistic GWAS methods based on our findings of **Chapter 3**, and to allow quantitative and binary traits to be analysed jointly. We identify novel genetic variants through multi-trait analyses, highlighting the utility of summary data for gaining further insight into the genetic aetiology of correlated traits.

In **Chapter 5** we further explore the utility of summary statistic data across multiple correlated phenotypes by building multivariate predictive models of MDD using PRS predictors in the UK Biobank, in addition to environmental (phenotypic) predictors. We build both main effect and interaction models, and use variable selection procedures for identifying the best fitting model. This study indicates that we can explain more variance in MDD case/control status by considering multiple predictors, and suggests that such an approach may lead to the identification of sub-types of heterogeneous disorders.

2. Multivariate simulation framework for genetic epidemiology

In this first chapter we develop a multivariate simulation framework suitable for genetic epidemiology, motivated by recent advances in multivariate genome-wide association study (GWAS) methodology. Although the standard approach to GWAS is to perform univariate analyses for each trait under study, the focus is switching to multi-trait analyses, facilitated by large national resources of data across a wide variety of phenotypes. The relative performance of current methodology, however, remains unclear; the aim of this chapter is to develop a simulation framework to act as a platform for the comparison of multi-trait GWAS methodology in order to elucidate their relative performance.

2.1 Introduction

The early stages of the GWAS era were dominated by studies with a single phenotype as outcome (Burton et al., 2007; The International Consortium for Blood Pressure, 2011; Teslovich et al., 2010), while in recent years multi-trait analyses have become popular (B. Bulik-Sullivan et al., 2015; Kauwe et al., 2014; Cross-Disorder Group of the Psychiatric Genomics Consortium and Genetic Risk Outcome of Psychosis (GROUP) Consortium, 2013). Multivariate methods have been developed to increase statistical power and identify pleiotropic loci in GWAS (Bottolo et al., 2013; Zhou and Stephens, 2012; O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Ferreira and Purcell, 2009; Klei et al., 2008; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007; Bolormaa et al., 2014; Casale et al., 2015; Huang et al., 2011; Zhang et al., 2014; Kim et al., 2016), while polygenic risk score and co-heritability estimation methods are now routinely applied to GWAS data to assess shared genetic aetiology across multiple traits (Purcell et al., 2009; Euesden et al., 2015; Dudbridge, 2013; Yang et al., 2011; B. K. Bulik-Sullivan et al., 2015; Krapohl et al., 2015; Vattikuti et al., 2012; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). However, these methods have been developed and applied in the absence of a dedicated simulation framework for generating data that reflects the complexity of multivariate data, where causal relationships are known.

Here we present a simulation framework designed to capture as much of the multivariate data landscape as possible, allowing researchers to benchmark methods across a range of simulation scenarios of genetic variants affecting multiple traits. We also incorporate real data so that simulated genetic effects and phenotypic correlations reflect reality. We provide a web application implementing our simulation

framework (www.MultiTraitGWAS.kcl.ac.uk). A built-in tool generates simple multivariate genetic data sets instantaneously, while a downloadable command-line program can be used to simulate larger, more complex multivariate data that can be used to test a range of multivariate methods under a variety of parameter settings. In **Chapter 3**, we use our simulation framework to perform a comparison study of the leading multi-trait GWAS methods (van der Sluis et al., 2013; O'Reilly et al., 2012; Zhu et al., 2015; Nath and Pavur, 1985; Ferreira and Purcell, 2009; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007). Finally, our web application provides a power calculator for multivariate GWAS, which should aid with optimal method selection given available data and in budgeting proposed studies. Our simulation framework and associated software should help to guide the future development and direction of multivariate methodology in genetic epidemiology.

2.2 Multivariate simulation framework

We construct a simulation framework to model the multivariate network that exists between a single nucleotide polymorphism (SNP) and K observed quantitative phenotypes (case/control phenotypes are modelled below), with internal and external risk factors and confounders (**Figure 5**). In **Chapter 3**, we exploit this simulation framework to test the relative performance of the leading multi-trait GWAS methods, and so this chapter acts as both a description of the study design for that comparison study and as an outline to our simulation software (www.MultiTraitGWAS.kcl.ac.uk) and its default settings, which can be changed according to user choice.

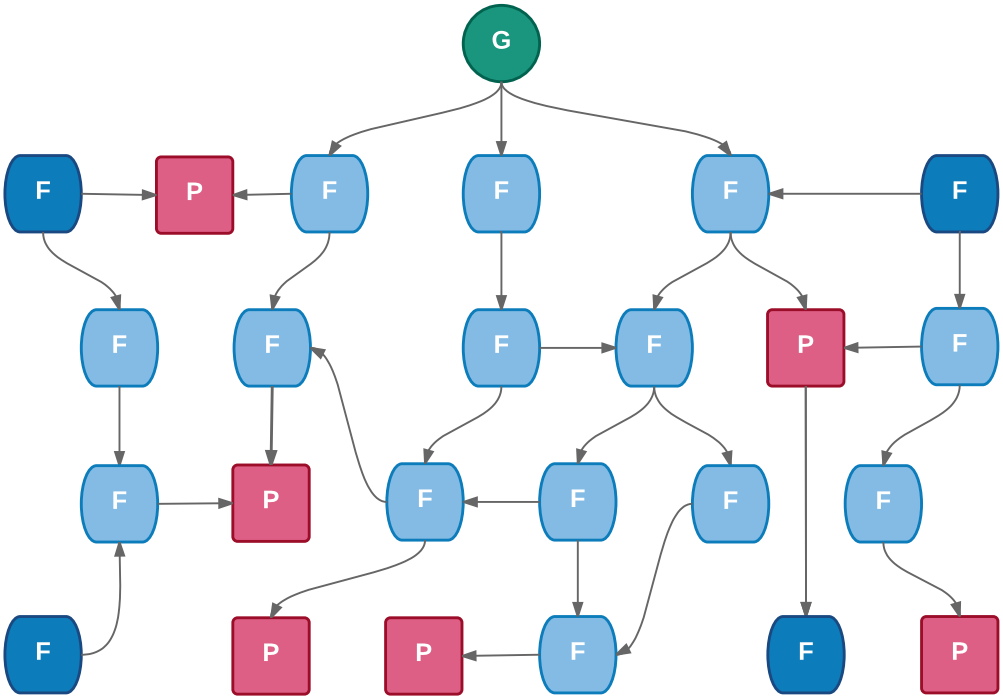

Figure 5. A biological network illustrating a genetic variant (G) influencing a set of biological entities, such as enzymes, metabolites and disease outcomes. Most are unmeasured internal (light blue) or external (dark blue) factors (F), but a subset corresponds to measured phenotypes to be tested (P).

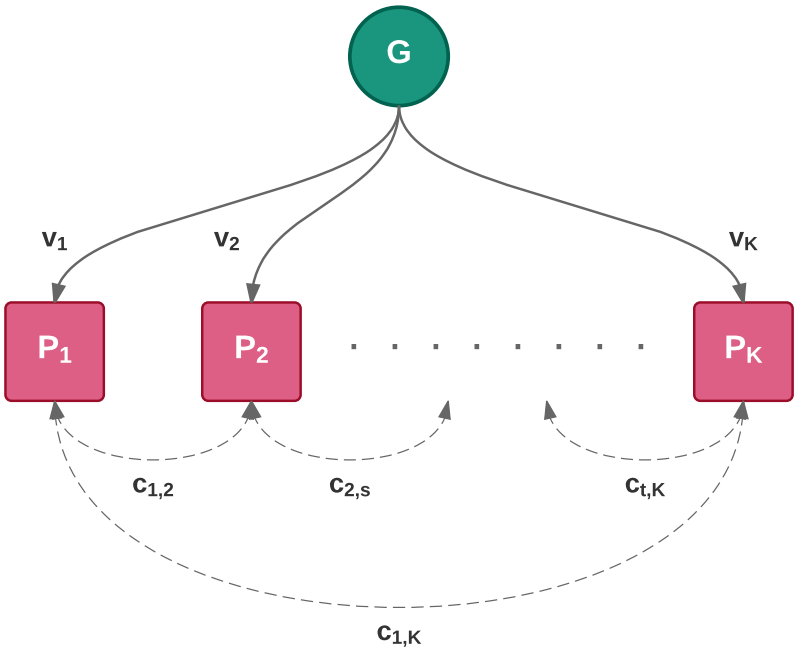

Figure 6. With no loss in generality, observed phenotype data from a biological network such as that represented in **Figure 5** (assuming no indirect genetic effects on observed phenotypes via other observed phenotypes) can be depicted and parameterised by $\boldsymbol{v}$ and $\boldsymbol{c}$ as shown. Values of $\boldsymbol{v}$ and $\boldsymbol{c}$ differ from their marginal values when observed risk factors are controlled for.

With no loss in generality in the context of multi-trait GWAS, this complex network collapses down to a simplified model consisting of a single genetic variant having direct effects on multiple correlated phenotypes (**Figure 6**), and can be modelled with two sets of parameters:

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ . \\ . \\ v_K \end{bmatrix}$$

$$\boldsymbol{c} = \begin{bmatrix} 1 & \cdots & c_{1,K} \\ \vdots & \ddots & \vdots \\ c_{1,K} & \cdots & 1 \end{bmatrix}$$

where $\boldsymbol{v}$ is the genetic effect vector of the variance in each of the K phenotypes explained by the genetic variant, and $\boldsymbol{c}$ models the phenotypic correlation matrix such

that $\boldsymbol{c}_{s,t}$ is the approximate (see below) correlation (Pearson's correlation coefficient, r) between phenotype s and phenotype t.

The SNP genotypes $G_i \in \{0, 1, 2\}$ are generated according to Hardy-Weinberg Equilibrium (HWE) in the proportions $\{p^2, 2pq, q^2\}$ where p is the major allele frequency and q is the minor allele frequency, with $q = 1 - p$. The genotype-phenotype data are simulated according to the model:

$$Y_i = f(\boldsymbol{v}) \cdot G_i + \varepsilon_i \quad \textbf{(1)}$$

where $Y_i = \{Y_{i,1}, \ldots, Y_{i,K}\}$ denotes the phenotype data corresponding to K phenotypes for an individual i, $f(\boldsymbol{v})$ denotes the regression coefficient corresponding to $\boldsymbol{v}$ phenotypic variance in Y_i explained by the SNP genotypes G_i, and ε_i is the residual variance drawn from the multivariate normal distribution $N(\mathbf{0}, \boldsymbol{c})$ (thus $\boldsymbol{c}$ is not the exact phenotypic correlation matrix but given small genotype effect sizes is approximately equivalent; henceforth we describe $\boldsymbol{c}$ as the phenotypic correlation matrix). The regression coefficient $f(\boldsymbol{v})$ is determined according to the following equation, under the assumption of additive genetic effects:

$$f(\boldsymbol{v}) = \sqrt{\frac{\tilde{v}}{2pq}}$$

where $\tilde{v}$ is the transformed phenotypic variance explained by the SNP, determined by:

$$\tilde{v} = \frac{v}{1 - v}$$

to ensure that the genetic variant explains $v\%$ of the phenotypic variance given that the residual variance is 1.

As default, we generate data corresponding to a sample of $5{,}000$ individuals and $10{,}000$ SNP replicates with minor allele frequency $(MAF) = 0.3$. While this, our main data-generating model, does not consider indirect effects of genetic variants on tested traits via other tested traits, nor case/control phenotype data, we also simulate and investigate these (see below).

By varying the values that $\boldsymbol{v}$ and $\boldsymbol{c}$ take, genotype-phenotype data consistent with a wide range of underlying biological networks (**Figure 5**) and set of observed phenotypes can be generated. However, since there are infinite values that $\boldsymbol{v}$ and $\boldsymbol{c}$ can take, a systematic search through the parameter space is required. Our simulation framework aims to capture as much of the parameter space as possible via four modelling scenarios:

- **(S1)** $\boldsymbol{v}$ and $\boldsymbol{c}$ are varied in a structured way
- **(S2)** $\boldsymbol{v}$ and $\boldsymbol{c}$ sampled from uniform distributions
- **(S3)** $\boldsymbol{v}$ and $\boldsymbol{c}$ reflect each other
- **(S4)** $\boldsymbol{v}$ and $\boldsymbol{c}$ informed by real data

While the simulation framework is sufficiently general that it should be useful outside of genetic epidemiology, we exploit it in **Chapter 3** to compare multi-trait GWAS methods, and thus the simulation scenarios chosen here reflect this.

Table 1 provides a summary of the multi-trait GWAS methods included in the comparison study. The methods are described in more detail in **Chapter 3**.

Method		Data Type	Information
Univariate	min-*P*	Univariate *P*-values	Adjusts univariate *P*-values in a standard Šidák correction, using the effective number of tests of Nyholt
	TATES	Univariate *P*-values	Adjusts univariate *P*-values using the *P*-value correlation matrix as determined by the phenotypic correlation matrix
Summary statistic	SHom	Univariate t-values	Performs a meta-analysis across traits weighted by the t-value correlations and the univariate study sample sizes
	SHet	Univariate t-values	Performs a test similar to SHom but for subsets of traits (determined by varying a threshold) and using absolute effect sizes
Individual-level genotype-phenotype data	MANOVA	Normally distributed	Test equivalent to multiple linear regression with phenotype predictors and genotype outcome
	CCA (mv-PLINK)	Normally distributed	Performs multiple linear regression with phenotype predictors and genotype outcome
	MultiPhen	Ordinal variable	Performs ordinal logistic regression with phenotype predictors and ordinal genotype outcome
	Combined-PC	Normally distributed	Performs a test equivalent to multiple linear regression with PC predictors and genotype outcome
	mv-BIMBAM	Normally distributed	Performs Bayesian multivariate regression, sub-setting the traits according to their SNP effect: direct, indirect or no effect
	mv-SNPTEST	Normally distributed	Performs Bayesian multivariate regression using a conjugate prior (Wishart on the covariance matrix, matrix normal on the genetic effects)

Table 1. Summary of the multi-trait GWAS methods included in the simulation framework.

2.3 Simulation scenarios

The simulation scenarios implemented in our simulation framework, and incorporated into our simulation software, are described below.

2.3.1 S1: Structured genetic effects and phenotypic correlations

In this scenario the genetic effects, $\boldsymbol{v}$, and phenotypic correlations, $\boldsymbol{c}$, are varied in a structured way. First we consider a case with only two phenotypes and three genetic effect vectors:

$$\boldsymbol{v_1} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$\boldsymbol{v_2} = \begin{bmatrix} 0.5 \\ 0.1 \end{bmatrix}$$

$$\boldsymbol{v_3} = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

such that $\boldsymbol{v_2}$ corresponds to a SNP that explains 0.5% variance in trait 1 and 0.1% variance in trait 2. The phenotypic correlations are varied between an r of -0.9 and 0.9 in increments of 0.1. Data corresponding to $5{,}000$ individuals are simulated according to **Equation 1** under each of the three effect vectors across the range of correlations. $10{,}000$ such replicates of genotype-phenotype data are simulated. Statistical power is measured as the proportion of results with a multivariate association *P*-value $< 5 \times 10^{-8}$ or a $\log_{10}$ Bayes factor > 6 for the Bayesian methods tested, corresponding, respectively, to the well established genome-wide significance *P*-value threshold and the Bayes factor for declaring a discovery proposed by Stephens and Balding (Stephens and Balding, 2009).

While the number of qualitatively different genetic effect vectors is only three for two phenotypes (equal effects, different effects, one effect and one with no effect), the number increases exponentially as more traits are considered. In scenario S1 we consider 2, 4, 8, 20 and 48 phenotypes, and for those with four or more we apply 10 genetic effect vectors, defined in **Table 2**. These effect vectors are chosen to cover a large proportion of qualitatively different combinations of genetic effects as efficiently as possible, in a systematic way, via assigning genetic effects to each $\frac{1}{4}$ of the traits.

Genetic effect vector	1st $\frac{1}{4}$ of traits	2nd $\frac{1}{4}$ of traits	3rd $\frac{1}{4}$ of traits	4th $\frac{1}{4}$ of traits
$\boldsymbol{v_1}$	0.5	0.5	0.5	0.5
$\boldsymbol{v_2}$	0.5	0.5	0.5	0
$\boldsymbol{v_3}$	0.5	0.5	0	0
$\boldsymbol{v_4}$	0.5	0	0	0
$\boldsymbol{v_5}$	0.5	0.5	0.5	0.1
$\boldsymbol{v_6}$	0.5	0.5	0.1	0.1
$\boldsymbol{v_7}$	0.5	0.1	0.1	0.1
$\boldsymbol{v_8}$	0.5	0.5	0.1	0
$\boldsymbol{v_9}$	0.5	0.1	0.1	0
$\boldsymbol{v_{10}}$	0.5	0.1	0	0

Table 2. Description of the 10 genetic effect vectors used in the simulations of 4 or more phenotypes. For 8 phenotypes, $\boldsymbol{v_5}$ corresponds to the genetic variant explaining 0.5% variance in 6 of the traits and 0.1% in 2 of the traits, while for 20 phenotypes $\boldsymbol{v_8}$ corresponds to the genetic variant explaining 0.5% variance in 10 traits, 0.1% variance in 5 traits and having no effect on 5 traits.

For two phenotypes we consider pairwise phenotypic correlations ranging between -0.9 and 0.9 in increments of 0.1. When we simulate four phenotypes, the correlation range is reduced to between -0.3 and 0.9, as the corresponding correlation matrices with pairwise phenotypic correlations less than -0.3 are not positive definite. For the same reason, the correlation range for eight phenotypes is between -0.1 and 0.9, and for 20 and 48 phenotypes the correlation range is between 0 and 0.9.

The simulations of scenario S1 consider only genetic effects in the same direction (i.e. positive genetic effects), although the phenotypic correlations can be negative, but in **Chapter 3** we explicitly consider the situation where one allele of the genetic variant increases the value of one trait but decreases the value of the other.

2.3.1.1 Modelling indirect effects

Our main data-generating model (**Equation 1**) does not consider indirect genetic effects, whereby the genetic variant has an effect on one of the tested phenotypes via its effect on another (a *downstream* or *mediated* effect). Here we perform simulations that model such an effect on two phenotypes. We simulate the genetic variant as explaining 0.5% variance in the first phenotype, and the first phenotype explaining 1%, 5%, 10% and 20% variance in the second phenotype.

We simulate and evaluate indirect genetic effects for scenario S1. We model an indirect genetic effect from a SNP G_i to a phenotype $Y_{i,2}$ by simulating a direct effect on a phenotype $Y_{i,1}$ and a direct effect from $Y_{i,1}$ to $Y_{i,2}$. Data are generated according to the following equations:

$$Y_{i,1} = f(\boldsymbol{v}) \cdot G_i + \varepsilon_{i,1}$$

$$Y_{i,2} = g(\boldsymbol{v}') \cdot Y_{i,1} + \varepsilon_{i,2}$$

where $f(\boldsymbol{v})$ denotes the regression coefficient corresponding to $\boldsymbol{v}$ phenotypic variance in $Y_{i,1}$ explained by the SNP, $g(\boldsymbol{v}')$ denotes the regression coefficient corresponding to $\boldsymbol{v}'$ phenotypic variance in $Y_{i,2}$ explained by $Y_{i,1}$, and $\varepsilon \sim \mathrm{N}(\mathbf{0}, \boldsymbol{c})$. We only simulate downstream effects for two phenotypes, but the simulations can be easily extended to incorporate more phenotypes, as well as more complex interaction networks.

2.3.1.2 Modelling case/control data

We simulate and evaluate case/control data for scenario S1. We first simulate quantitative phenotype data as in **Equation 1**, and then apply a liability threshold model (Falconer, 1960) of disease to generate case/control phenotype data according to the prevalence of the disease:

$$Y_{i,k} = \begin{cases} 1 & Y_{i,k} \geq qnorm(1 - prev_k) \\ 0 & Y_{i,k} < qnorm(1 - prev_k) \end{cases}$$

where $prev_k$ is the prevalence of the k^{th} phenotype. In **Chapter 3**, simulations of two case/control phenotypes are performed, as well as a mixture of one case/control phenotype and one quantitative trait.

2.3.2 S2: Uniform genetic effects and phenotypic correlations

In contrast to the structure of the first simulation scenario, here we choose genetic effects and phenotypic correlations uniformly, where every value is assumed to be equally likely. The instances of genetic effects and phenotypic correlations simulated in the first scenario are highly structured, which allows us to more easily infer continuous power curves across the full correlation range for given genetic effects, than from an unrelated set of point estimates in high-dimensional model space. However, this may result in simulating unrealistic data, such as a SNP explaining 0.5% variance in 10 traits that all have pairwise correlations of 0.1. By sampling parameters uniformly, we eradicate the structure in our choices and allow the parameters to take any values within a prescribed range. In this way, the power estimates we achieve in this scenario should act as an average of other more structured scenarios, and provide a more general indication of the performance of the methods.

We sample the genetic effects from the uniform distribution:

$$U(0, 0.005)$$

such that the genetic effects can take any value between 0% variance explained and 0.5% variance explained. We sample the pairwise phenotypic correlations from the uniform distribution:

$$U(-0.9, 0.9)$$

such that the pairwise phenotypic correlations can take any value between -0.9 and 0.9. We perform this sampling for each pair of phenotypes, and construct the corresponding phenotypic correlation matrix. This matrix is required to be positive definite, and so the matrix is resampled if the previous sampling of pairwise phenotypic correlations did not result in a positive definite matrix, but otherwise each of the $10{,}000$ genotype-phenotype simulated datasets relate to a random combination of genetic effects and phenotypic correlations.

While this scenario simulates a random and diverse set of effects and correlations and thus may be vulnerable to simulating unrealistic combinations of genetic effects and phenotypic correlations, the aim here is to generate unrestricted parameter values in contrast to the first scenario; by doing so, we hope to gain an overall picture of the performance of multi-trait methods rather than that relating only to specific scenarios.

2.3.3 S3: Genetic effects reflective of phenotypic correlations

Since most causal genetic variants explain $< 1\%$ of phenotypic variance (Park et al., 2010; Visscher et al., 2012), their effects on a set of phenotypes do not induce phenotypic correlations reflecting their relative sizes. However, it may be likely that genetic effects are on average more reflective of the corresponding phenotypic correlations than not. Here we simulate data such that the phenotypic correlations reflect the relative sizes of the genetic effects. Phenotypic correlations are chosen to reflect the genetic effect vectors described in scenario S1 (three effect vectors for two traits, and 10 for four or more traits) in the following way: if the variant has equal effects on two traits then the corresponding phenotypic correlation is set to be 0.6, for different effect sizes the correlation is 0.2, and if the variant affects one phenotype but not the other then their pairwise correlation is set as 0.05.

In this way, we align the phenotypic and genetic correlations so that they are concordant, which leads to simulation of potentially more realistic genotype-phenotype data and avoids modelling scenarios that are highly unlikely to occur in biological data. As in scenario S1, the power of the methods will be a function of the phenotypic correlations modelled, the impact of which can be further explored using our simulation tool.

2.3.4 S4: Real data informed genetic effects and phenotypic correlations

The final simulation scenario exploits real data from published GWAS results, and from the individual-level genotype-phenotype data of the Northern Finland Birth

Cohort 1966 (NFBC1966) on $4{,}772$ individuals, to simulate realistic values for the genetic effect and phenotype correlation parameters. This scenario is in two parts:

(a) Real data informed phenotype correlations

(b) Real data informed genetic effects and phenotype correlations

2.3.4.1 (a) Real data informed phenotype correlations

Here we fit a mixture Gaussian distribution to the observed Northern Finland Birth Cohort 1966 (NFBC1966) phenotype correlations of 16 metabolic traits. We used this set of metabolic traits, which include BMI, systolic and diastolic blood pressure, HDL and LDL, because they represent a large number of well-measured quantitative phenotypes with diverse pairwise correlations. We fit a theoretical distribution to this in order to sample uniquely from it repeatedly; the fitted mixture Gaussian distribution is given by:

$$\frac{1}{20}\,\mathrm{N}(-0.23, 0.045^2) + \frac{9}{10}\mathrm{N}(0.21, 0.175^2) + \frac{1}{20}\mathrm{N}(0.74, 0.07^2) \qquad \textbf{(2)}$$

The observed NFBC1966 and fitted probability density functions are shown in **Figure 7**. Genotype-phenotype data are generated as in scenario S1 but by sampling the pairwise phenotypic correlations from this fitted density, discarding sampled non-positive definite correlation matrices. This scenario is similar to scenarios S1 and S2 in that the genetic effect vectors are independent of the phenotypic correlations. However, the phenotypic correlations are based on real data here and so these data should be more reflective of epidemiological data.

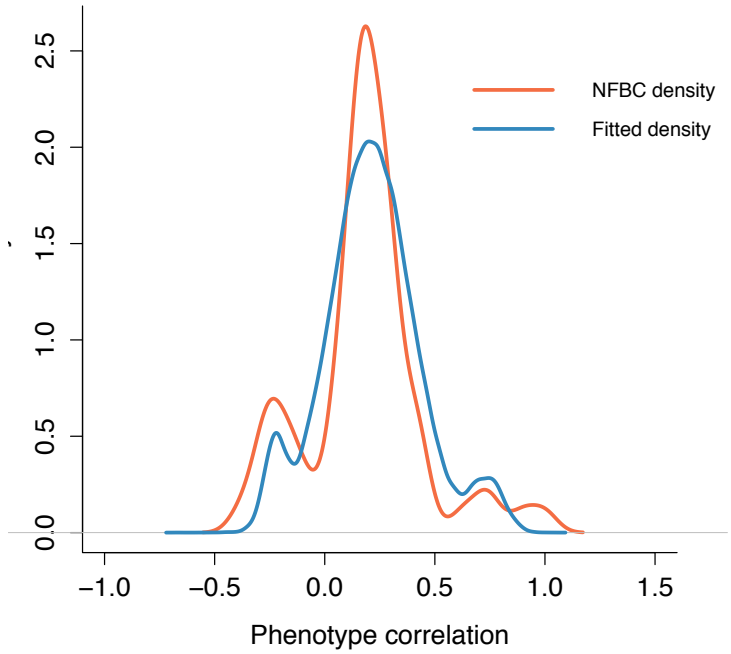

Figure 7. Phenotypic correlation density based on 16 metabolic traits from the NFBC1966, and fitted mixture Gaussian density as given in **Equation 2**. Pairwise phenotypic correlations are sampled from this fitted density for simulations of scenario S4a.

2.3.4.2 (b) Real data informed genetic effects and phenotype correlations

Here we sample genetic effect sizes directly from reported genotype-phenotype associations from publicly released GWAS on 12 phenotypes: height (Wood et al., 2014), BMI (Yang et al., 2012), systolic and diastolic blood pressure (The International Consortium for Blood Pressure, 2011), Triglycerides, HDL, LDL and total cholesterol (Global Lipids Genetics Consortium, 2013), fasting-glucose, fasting-insulin, HOMA-B and HOMA-IR (Dupuis et al., 2010). We compiled a list of all SNPs with a reported genome-wide significant association in the largest available GWAS on each trait, and recorded the corresponding genetic effect size for each SNP across all 12 traits. Across the 12 phenotypes under study there are a total of 237 unique SNP-phenotype associations with complete summary data across all traits, and thus 237 SNPs eligible for inclusion in the multi-trait analyses. From this set of SNPs, for traits with 20 or fewer genome-wide significant associations, all associated SNPs are taken forward for analysis. Of the remaining traits, 20 SNPs are sampled

from the larger set of associated SNPs in order to reduce bias toward traits with a larger number of genome-wide significant findings. This approach is applied for the analysis of all 12 phenotypes, as well as for the subsets of 2, 4 and 8 phenotypes.

We use the β (SNP effect size) estimates from these published GWAS to inform the simulation of genetic effects. For a given SNP G_i we take the absolute effect size estimates across K phenotypes, say $\beta_1, \ldots, \beta_K$, and apply a transformation so that the maximum effect size is 0.5% of phenotypic variance, while maintaining the relative effect sizes. The transformation factor, $d_{i,}$ for SNP G_i is defined as follows:

$$d_i = \frac{\beta_S}{\beta_i^M}$$

where

$$\beta_S = \sqrt{\frac{0.005}{2pq}}$$

where $p = 1 - MAF$, $q = 1 - p$, β_S is the beta coefficient that corresponds to the maximum effect size of 0.5% variance explained, and β_i^M is the maximum beta for a given SNP G_i across all K phenotypes. The real data obtained beta coefficients for SNP G_i are then multiplied by d_i to generate the beta coefficients used to simulate the phenotype data according to **Equation 1**.

The real effect sizes are transformed so that the largest effect is equivalent to 0.5% of phenotypic variance so that the results of scenario S4b can be directly compared to those of the other scenarios, since power is a function of effect size. The relative effect sizes of the set of SNP-phenotype associations remain the same after the transformation.

We then simulate $10{,}000$ replicates of genotype-phenotype data corresponding to $5{,}000$ individuals by sampling from these genetic effect vectors and directly (not from the fitted density of **Figure 7**) from the phenotypic correlations estimated in the NFBC1966 data on those traits. As well as simulations based on all 12 phenotypes, we repeat this scenario for 2, 4 and 8 phenotypes by iterating through all ${}^{12}C_K$ combinations of the 12 phenotypes, forming the corresponding genetic effect vectors and phenotype correlation matrices to simulate the data.

2.4 Comparison with previous multivariate genetic simulations

We designed the simulation settings implemented in this framework to be as comprehensive as possible by considering a variety of different simulation scenarios, including those that likely closely match reality. Previous simulation efforts, while seemingly considering a range of scenarios, have not fully explored the available model space and have thus provided an incomplete picture of the performance of multi-trait methods.

2.4.1 van der Sluis *et al.* 2013: TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies

Publications of novel methodology are often accompanied by simulation studies showing that the novel method is preferable to existing methodology in terms of statistical power. An example is the publication of the TATES method (van der Sluis et al., 2013), which produces an omnibus *P*-value for the association of each SNP with a group of phenotypes from a set of corresponding univariate *P*-values from published GWAS. The authors considered a series of different network models to

assess the performance of their method. One type of model considered is a single factor network where a genetic variant has an effect on a single factor, which then affects a series of traits (**Figure 8a**), or the genetic variant effects a trait directly, which is correlated with other traits via a single factor (**Figure 8b**).

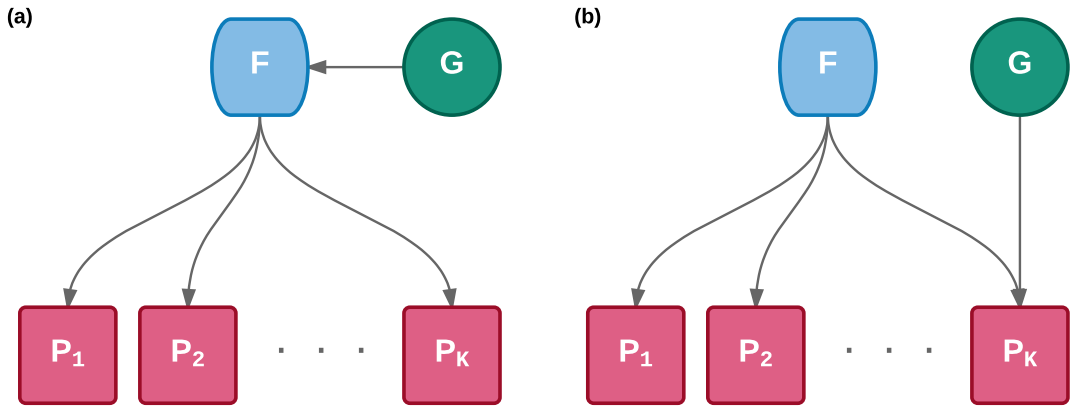

Figure 8. Single factor biological networks where **(a)** the genetic variant has a direct effect on a mediating factor, which has effects on multiple phenotypes, and **(b)** the genetic variant has a direct effect on one phenotype, which is correlated with other modelled phenotypes via a single factor.

Multi-factorial models are also considered where the genetic variant has an effect on a single factor, which then affects a series of traits, with a correlation between this factor and other factor(s) (**Figure 9a**), or the genetic variant affects a trait directly, which is correlated with other traits via numerous correlated factors (**Figure 9b**).

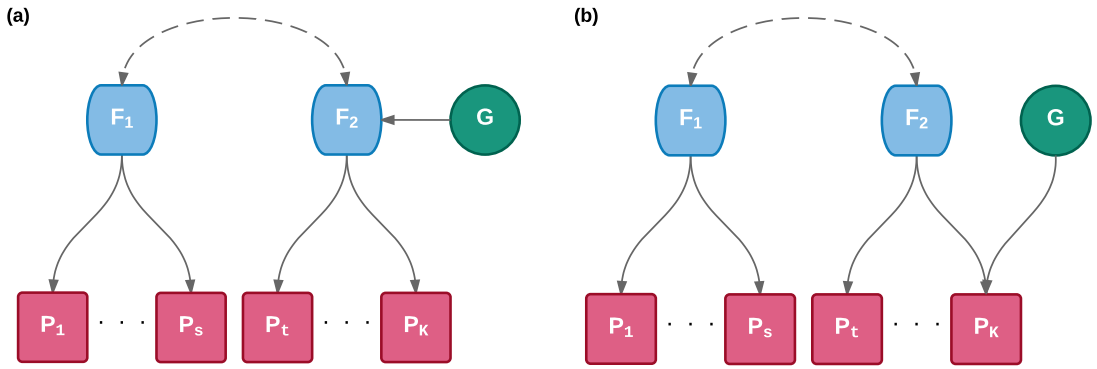

Figure 9. Multi-factorial biological networks where **(a)** the genetic variant has a direct effect on a mediating factor, which is correlated with another factor, both of which have effects on phenotypes, and **(b)** the genetic variant has a direct effect on one phenotype, which has a correlation structure with the other modelled phenotypes via the two factors.

The single factor models correspond directly to the model used in our simulation framework to simulate direct genetic effects. Simulations of the factor model given by **Figure 8a** can be achieved by our simulation framework; the impact of the genetic effects going via a mediating factor are that the genetic effects on the traits are diminished. However, by choosing appropriate values for the genetic effect from the variant to the factor and from the factor to the traits, the whole model space can be explored using our framework. Here the phenotypic correlation is induced by the mediating factor, but in our model we can stipulate the phenotypic correlations explicitly. The factor model of **Figure 8b** corresponds to a special case of the one factor model, where the genetic variant affects only one trait, and is correlated with other traits via some factor. In our simulation framework the two one-factor models are considered to be the same modelling scenario where we can achieve the model of **Figure 8a** by simulating genetic effects on all correlated traits, and **Figure 8b** by simulating only one trait to be affected by the genetic variant, and inducing a phenotypic correlation between the traits.

The modelling scenarios of the multi-factorial models can also be achieved using our simulation framework via appropriate choice of the genetic effects and phenotypic

correlations. In both **Figure 9a** and **Figure 9b,** the two factor models induce a correlation structure in the phenotypes due to the two groupings of factors and phenotypes, which can be achieved in our simulation framework by appropriate choice of the phenotypic correlation matrix i.e. by having sub-matrices that correspond to the two groups of phenotypes. In **Figure 9a** the genetic variant (G) is affecting one factor (F_2), which is correlated with the other factor (F_1). This means that the genetic effect on traits P_t to P_K is mediated by F_2 and thus diminished, while the genetic effects on phenotypes P_1 to P_s are further diminished according to the correlation between the two factors. Like in **Figure 8,** the scenario represented by **Figure 9b** is a special case of the two-factor model of **Figure 9a**, where only one trait is affected by the genetic variant, and there exists a correlation structure between the traits. Again, in our simulations these are considered to be one modelling scenario, where both **Figure 9a** and **Figure 9b** can be achieved by choosing the genetic effects appropriately, and setting up the phenotypic correlation matrix to induce the correlation structure of a two-factor model.

Finally, a network model is considered where the genetic variant affects a trait, with downstream effects occurring between a larger number of traits in the interconnected network.

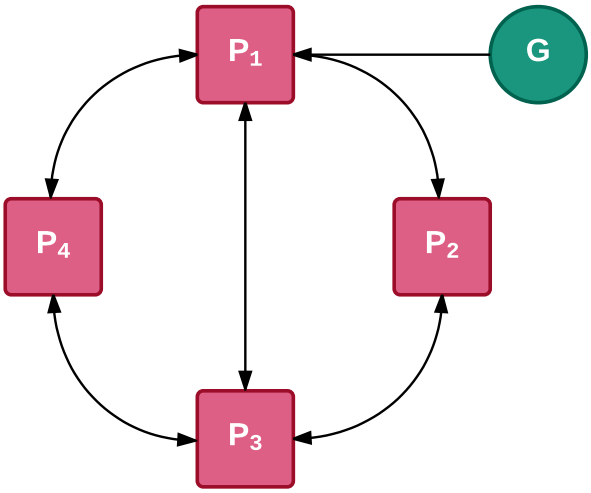

Figure 10. Network model where a genetic variant (G) affects a single phenotype P_1 which is part of an interconnected network of phenotypes with effects between phenotypes.

The network model of **Figure 10** is the most different to the other models considered. Here a genetic variant (G) affects a single phenotype P_1, which is in a larger network of phenotypes, where there exists not only correlations between the phenotypes but also causal effects between them. This type of model can be achieved in our downstream analyses of scenario S1, where downstream effects between traits exist. While we only directly consider two traits for the purposes of our comparison, more traits could easily be incorporated.

In the context of assessing the performance of multi-trait GWAS methods, the key factors are the sample size, the magnitude of genetic effects on the traits being analysed, and how the traits are correlated, meaning that the results pertaining to these various factor models can be obtained by varying the genetic effects and phenotypic correlations in our framework appropriately. By varying these two sets of parameters systematically across a wide range of possible values we can gain insight into the performance of multi-trait GWAS methods across different modelling scenarios relevant to genetic epidemiology.

In the paper presenting the TATES method (van der Sluis et al., 2013) a mixture of different sizes of genetic effect is not considered in these factor and network models. Although data are simulated for genetic effects between 0% and 1% variance explained in increments of 0.1, varying the effect size equally across all traits is equivalent to simply varying the sample size, and there is no exploration of the effect of mixed genetic effects. Continuous, binary and ordinal traits are simulated and methods are applied to a mixture of different types of phenotype, as well as to just continuous traits, with up to 20 traits simulated. However, the correlations of these traits – or correlations between different groupings of the traits in the case of the multi-factorial models – are not varied systematically, rather specific point values of the correlations are chosen and these chosen correlations are all positive.

2.4.2 Galesloot *et al.* 2014: A Comparison of Multivariate Genome-Wide Association Methods

Galesloot and colleagues (Galesloot et al., 2014) performed a comparison study of multi-trait GWAS methods using simulated data on just three traits, with three scenarios of genetic effect: the genetic variant affects only one trait, the genetic variant affects two traits, and the genetic variant affects all three traits (genetic effects fixed at 0.1%). While these genetic effects will capture a pleiotropic effect of the genetic variant on the traits, as well as only a single trait being affected, a mixture of genetic effect sizes is not considered here, and the small number of traits analysed limits the model space dramatically. Only three levels of phenotypic correlation $(0, 0.3$ and $0.7)$ were simulated rather than exploring the full correlation range, meaning that only a snapshot of the performance of the methods is provided, and no negative phenotypic correlations are considered. The other parameter that is varied is the minor allele frequency (MAF). However, this corresponds to varying the

effect size in terms of statistical power, and so does not provide additional insight into the relative performance of the methodology and does not make clear by how much the effect size has changed in terms of phenotypic variance explained.

Positive and negative genetic correlations are also considered, where there exists a positive correlation when the genetic effects are in the same direction, and a negative correlation when the genetic effects are in opposite directions. This involves the determination of the sign of the beta coefficient of the SNP effect on the traits. If, for two traits, the signs are the same then this will induce a positive genetic correlation; if the signs are opposite then the traits will be negatively genetically correlated. We would expect the results obtained by simulating negative genetic correlations to correspond to those with a positive genetic correlation but where a negative phenotypic correlation is simulated, which we consider in our simulations of scenario S1, the results of which are presented in **Chapter 3**.

While the simulation scenarios used in this comparison study did provide some indication of the performance of multi-trait GWAS methods, they only captured a limited snapshot of the possible modelling scenarios, and so we cannot construct the full picture of the performance of the methods from these results alone. Other important factors for the performance of multi-trait GWAS methods were not considered, such as modelling large numbers of traits, a mixture in genetic effect sizes and freedom in the sampling of the phenotypic correlations.

2.5 Discussion

We have presented a simulation framework for generating data relating genetic variants with multiple phenotypes. The framework incorporates a range of simulation scenarios to explore the vast model space relevant to multivariate genetic data, thus providing a consistent platform for benchmarking multivariate methods, of sufficient rigor to expose their differences and similarities and to demystify user choice.

The motivation behind this simulation framework was to construct simulation scenarios relevant to multi-trait GWAS methods, in order to facilitate a comprehensive comparison between them. We focus on the modelling challenges of single-SNP, multi-trait methods, but there are methods that also model multiple SNPs jointly (Bottolo et al., 2013; Zhou and Stephens, 2012; Kim et al., 2016). Modifications could be made to the simulation framework presented here to make it applicable to this type of method, by generating correlated SNP data and simulating groups of SNPs with effects on sets of correlated phenotypes. While we consider a wide range of possible modelling scenarios, due to the infinite combinations of genetic effect and phenotypic correlations, we inevitably had to restrict to a feasible, yet broad, set of parameters. However, our simulation software will facilitate the exploration of many more simulation scenarios by varying the simulated causal relationships between genetic variants and correlated sets of phenotypes.

While we go on to exploit the simulation framework for a multi-trait GWAS methods comparison, our framework and accompanying simulator should have wide application across genetic epidemiology. Moreover, with minor modifications, such as modelling a normally distributed risk factor such as a polygenic risk score (PRS), the simulation framework could be exploited to model any network of correlated variables

for which a subset are influenced by a common factor. Thus our simulation framework, implemented as a web application and open-source software program (www.MultiTraitGWAS.kcl.ac.uk) has potential utility across the life sciences, economics and any discipline involving correlated variables.

3. Comparison and investigation of the performance of multi-trait GWAS methods

In **Chapter 2** we presented the multivariate simulation framework that forms the basis and motivation for the analyses in this chapter. While our simulation framework and associated software will be useful for multivariate methodology development and applications across genetic epidemiology, we exploit it here specifically to perform a comprehensive comparison of multi-trait GWAS methods and to investigate the features of their performance.

3.1 Introduction

Burgeoning availability of genome-wide association study (GWAS) results and national biobank data has led to growing interest in performing multi-trait genetic analyses. Numerous multi-trait GWAS methods that exploit either summary statistics or individual-level data have been developed (O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Ferreira and Purcell, 2009; Klei et al., 2008; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007; Bolormaa et al., 2014; Casale et al., 2015; Zhang et al., 2014; Kim et al., 2016), but their relative performance is unclear. Here we exploit the simulation framework presented in **Chapter 2** to perform a comprehensive comparison of the leading single-SNP multi-trait methods, both those that use individual-level genotype-phenotype data: MANOVA (Nath and Pavur, 1985), CCA (mv-PLINK) (Ferreira and Purcell, 2009), Combined-PC (Aschard et al., 2014), MultiPhen (O'Reilly et al., 2012), mv-BIMBAM (Stephens, 2013) and mv-SNPTEST (Marchini et al., 2007), and those that exploit GWAS summary statistics: min-*P* (O'Reilly et al., 2012), TATES (van der Sluis et al., 2013), S_{Hom} and S_{Het} (Zhu et al., 2015). These methods cover several approaches to testing the association of genetic variants with multiple phenotypes, including multiple linear regression techniques (Ferreira and Purcell, 2009; Aschard et al., 2014), a reversed (ordinal) regression with SNP as outcome (O'Reilly et al., 2012), simple (O'Reilly et al., 2012) and complex (van der Sluis et al., 2013) adjustments of summary statistic results, across trait meta-analysis techniques (Zhu et al., 2015) and Bayesian methods (Stephens, 2013; Marchini et al., 2007).

The simulation framework was developed to provide a thorough and consistent platform on which to compare and benchmark these different methods. We illustrate statistical power across a variety of combinations of genetic effects and phenotype

correlations for up to 48 phenotypes, which we believe represents the clearest way to expose differences in method performance. In addition, we provide insight into the behaviour of the different approaches to multi-trait GWAS, and shed light on when methods should be implemented in order to optimise discovery power. Our findings provide the clearest picture to date of the relative performance of multi-trait GWAS methods and act as a guide for method selection in the field.

3.2 Material and Methods

3.2.1 Multivariate simulation framework

The simulation framework utilised in this comparison of the leading multi-trait GWAS methods was presented in **Chapter 2**. **Table 3** provides a summary of the simulation scenarios considered.

Scenario	Phenotypes	Genetic Effects	Phenotypic Correlations
S1	$2, 4, 8, 20$ and 48	Fixed genetic effects defined by $\boldsymbol{v_1} - \boldsymbol{v_3}$ for 2 traits and **Table 2** for 4 or more traits	Pairwise phenotypic correlations range between -0.9 and 0.9 in increments of 0.1
S2	$2, 4$ and 8	Genetic effects sampled uniformly between 0% and 0.5% phenotypic variance explained	Phenotypic correlations sampled uniformly, ensuring the resulting correlation matrix is positive definite
S3	$2, 4, 8, 20$ and 48	Genetic effects chosen to reflect the phenotypic correlations	Pairwise phenotypic correlations chosen according to the genetic effects
S4a	$2, 4$ and 8	Fixed genetic effects defined by $\boldsymbol{v_1} - \boldsymbol{v_3}$ for 2 traits and **Table 2** for 4 or more traits	Phenotypic correlations sampled from a fitted mixture Gaussian distribution based on NFBC1966 data
S4b	$2, 4, 8$ and 12	Genetic effects based on univariate GWAS summary statistics	Phenotypic correlations obtained directly from the NFBC1966 on the phenotypes for which the genetic effects were obtained

Table 3. Summary of the simulation scenarios that comprise the simulation framework presented in **Chapter 2**, and used here to perform a comparison of multi-trait GWAS methods.

3.2.2 Multi-trait GWAS methods

The 10 methods included in the comparison study are briefly described below. These methods were chosen due to a combination of them being highly cited and representing a diverse set of approaches to multi-trait GWAS. While methods exist that simultaneously model multiple SNPs and multiple traits (Bottolo et al., 2013; Zhou and Stephens, 2012; Kim et al., 2016) we focus on the more common single-SNP methods here to isolate the methodological advances responsible for the greatest increases in power when modelling multiple phenotypes.

3.2.2.1 Univariate adjustments

We include two methods in our comparison study that perform adjustments to univariate *P*-values to obtain a joint-association *P*-value across traits. These methods do not, however, explicitly model the phenotypes jointly. The minimum joint *P*-value for a SNP obtained by these methods is determined by the minimum *P*-value at that SNP across all traits under study. Thus, these methods cannot identify novel findings, but are included here as a comparison to the univariate approach.

min-*P*: This test was proposed in O'Reilly *et al.* (O'Reilly et al., 2012) as a way of comparing MultiPhen to a simple multi-trait approach that exploits only existing GWAS univariate summary statistics. First, the minimum *P*-value from the group of K *P*-values corresponding to the K phenotypes under study is recorded for every SNP, using the published univariate GWAS results. Next the effective number of independent tests represented by the results on the K phenotypes is estimated using the correlation matrix of the phenotypes according to Nyholt (Nyholt, 2004), and then the recorded minimum *P*-value is adjusted according to this number of tests in a standard Šidák correction (Sidak, 1971).

TATES: This test is similar to that of min-*P* but performs a more sophisticated correction for multiple testing across the different phenotype results (van der Sluis et al., 2013). Here the results are ranked according to *P*-value and then the extended SIMES procedure of Li *et al.* (Li et al., 2011) is performed – on multiple traits rather than variants – by progressively re-ordering the minimum *P*-value according to a scaling that is a function of the effective number of independent *P*-values.

3.2.2.2 Summary statistic GWAS methods

We include two methods that use summary statistics (beta and standard error estimates) from univariate GWAS to perform multi-trait analyses. These methods perform meta-analyses across traits and cohorts.

S_{Hom}: This test combines Wald test statistics from univariate GWAS summary statistics relating to a SNP across both multiple cohorts and multiple phenotypes in a meta-analysis (Zhu et al., 2015). Heterogeneity in effect size and statistical power across cohorts is accounted for, as is the correlation among the test statistics, while the overall test statistic has optimal power when the genetic effect is homogeneous across traits and cohorts. Of the 10 tests considered here, S_{Hom} is that which can be most considered a 'test for pleiotropy', being equivalent to a meta-analysis of effect sizes across traits and cohorts with optimal power under fixed effects.

S_{Het}: This test is derived from S_{Hom} but is designed to detect genetic variants that only affect a subset of the total number of traits under study (Zhu et al., 2015). Only those traits with a corresponding Wald test statistic above some threshold are included in the calculation of a statistic equivalent to that of S_{Hom}. This is then recalculated across the range of possible threshold values with the maximum value obtained being the test statistic S_{Het}. Since this S_{Het} statistic does not follow a standard theoretical distribution, *P*-values are computed via simulation.

3.2.2.3 Individual-level GWAS methods

Included in our comparison study are a range of different approaches to performing multi-trait GWAS that exploit individual-level genotype-phenotype data. The development of individual-level methods has been an active area of research in genetic epidemiology, which is reflected in the diverse approaches developed here.

MANOVA: The standard Multivariate Analysis of Variance statistical test (Nath and Pavur, 1985), which is the multivariate extension of ANOVA, and equivalent to a reversed multivariate linear regression with genetic variant as outcome (O'Reilly et al., 2012).

CCA (mv-PLINK): Canonical Correlation Analysis (CCA) refers to a statistical procedure for identifying and testing the association of linear combinations of two sets of variables that maximise their correlation. While this approach could theoretically be applied to test for association between multiple genetic variants and traits jointly, in the context of multi-trait, single-SNP analyses, as incorporated into PLINK (Ferreira and Purcell, 2009), this test is equivalent to a reversed multivariate linear regression with a single genetic variant as outcome (O'Reilly et al., 2012). This method is equivalent to MANOVA, and has been shown to be approximately equivalent to MultiPhen (see below) for common SNPs (O'Reilly et al., 2012).

Combined-PC: This test performs a principal components analysis (PCA) on the phenotype data (Aschard et al., 2014). Separate univariate linear regressions are performed each with a different PC as outcome and SNP as predictor, and then the chi-squared statistics corresponding to the SNP-PC association from each regression are summed. Since the PCs are orthogonal to each other these tests are independent and their results can thus be summed in this way. Given small genetic effect sizes, the simple linear regressions of PC on SNP are approximately

equivalent to reverse regressions of SNP on PC; the sum of these individual regressions is then equivalent to a multiple linear regression with PC predictors. Since all PCs are included as predictors, this is equivalent to a multiple linear regression with phenotype predictors and SNP as outcome, and thus overall the Combined-PC method is approximately equivalent to CCA (mv-PLINK) (Ferreira and Purcell, 2009).

MultiPhen: This test performs a 'reversed regression', with multiple phenotype predictors and genetic variant as outcome (O'Reilly et al., 2012). Since genotypes of SNPs (and other genetic variants) correspond to ordinal data, an ordinal regression is performed here. This test has been shown to be equivalent to MANOVA and CCA for common SNPs (O'Reilly et al., 2012).

mv-BIMBAM: This test (Stephens, 2013) performs Bayesian multivariate regression by partitioning the phenotypes according to the effect of the genetic variant on them: direct, indirect or no effect. Statistical power is assessed using a $\log_{10}$ Bayes factor threshold of 6, following Stephens and Balding (Stephens and Balding, 2009) and Shim *et al.* (Shim et al., 2015).

mv-SNPTEST: This test performs Bayesian multivariate regression using an inverse Wishart distribution and matrix normal priors (Marchini et al., 2007). The fit of the full model is compared to that of the null model, and a $\log_{10}$ Bayes factor quantifies the association between SNP and phenotypes. Statistical power is assessed using a $\log_{10}$ Bayes factor threshold of 6 (Stephens and Balding, 2009).

3.3 Multi-trait GWAS method comparison study

The illustration of method performance in the literature is often challenging to interpret and is highly inconsistent across publications. Researchers are thus left with a perplexing choice between competing methods. Our dedicated simulation framework enables a systematic and rigorous search across the multivariate model space. We present results across a range of genetic effects and phenotypic correlations, from which a clear picture of the relative performance of the methods emerges. Our findings can guide the design of future GWAS, in particular those utilising the rich multivariate data becoming available from large-scale biobanks such as the UK Biobank, German National Cohort and planned US Biobank.

3.2.3 S1: Structured genetic effects and phenotypic correlations

In this scenario the genetic effects, $\boldsymbol{v}$, and phenotypic correlations, $\boldsymbol{c}$, are varied in a structured way; for simulations of two phenotypes we consider three genetic effect vectors:

$$\boldsymbol{v_1} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$\boldsymbol{v_2} = \begin{bmatrix} 0.5 \\ 0.1 \end{bmatrix}$$

$$\boldsymbol{v_3} = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

The phenotypic correlations are varied between -0.9 and 0.9 in increments of 0.1. **Figure 11** shows the statistical power of the 10 multivariate methods when applied to these data.

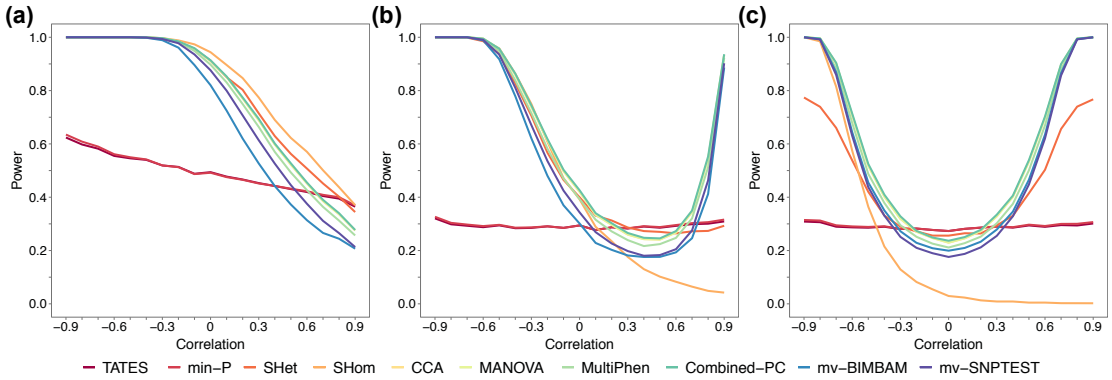

Figure 11. (a) The genetic variant explains 0.5% variance in two traits ($\boldsymbol{v_1}$). **(b)** The genetic variant explains 0.5% variance in one trait and 0.1% in the other ($\boldsymbol{v_2}$). **(c)** The genetic variant explains 0.5% variance in one trait and has no effect on the other ($\boldsymbol{v_3}$).

Figure 11 shows that the methods fall in to one of two distinct groups in terms of their power curves, except for S_{Hom}, which has a different pattern in **Figure 11b** and **Figure 11c**. The min-*P* and TATES methods – which have almost identical power – have lower power across much of the parameter space. In **Figure 11a**, we observe a decrease in power for min-*P* and TATES across the correlation range. Both methods perform univariate tests for association with each trait before correcting the smallest *P*-value to account for the number of tests performed. When the traits are highly positively correlated the variability in the univariate *P*-values is small, resulting in similar *P*-values for each test. In contrast, when the traits are uncorrelated the two tests are independent and the variability in the univariate *P*-values is greater, increasing the probability that one is small. For negative phenotypic correlations, this variability is even greater, resulting in higher power. However, in **Figure 11b** and **Figure 11c**, since the genetic effect on the second trait is very small or zero, then the minimum *P*-value will almost certainly derive from the SNP with large effect and is thus invariant to the phenotypic correlation.

The group of methods that follow a different pattern gain power via reducing the residual variance in analysing the traits jointly. When the genetic variant affects both traits equally and the traits are highly correlated (**Figure 11a**), these methods lose power due to the limited additional residual variance explained for the same degrees of freedom penalty. When only one trait is affected by the genetic variant and the phenotypes are uncorrelated (**Figure 11c**) then there is no gain in residual variance explained by including the unaffected trait. The methods gain power when there is discordance between the genetic effects and the phenotypic correlations, due to the potential increase in explained residual variance. **Figure 11c** shows that while the S_{Het} method follows the same pattern as the other methods in this group, it has lower power for high positive and negative correlations. This is due to the trait sub-setting procedure of S_{Het}, which incurs a relatively strong, multiple testing penalty. However, as a summary statistic method S_{Het} can be applied to large publicly available GWAS results and thus may have a substantial boost in power for certain traits.

In **Figure 11** the S_{Hom} method, which performs a meta-analysis on the traits, generally performs best in pleiotropic scenarios. In **Figure 11c** only one trait is affected by the genetic variant and when the traits are uncorrelated there is no gain in power by including the unaffected trait; S_{Hom} loses power over min-*P* and TATES here due to it relating to the average rather than maximum association. For highly positively correlated traits, the addition of the unaffected trait in the meta-analysis reduces the average effect size and thus power, but for highly negatively correlated traits the individual effect sizes are augmented, leading to increased power. While a standard meta-analysis would produce a different power curve to that of S_{Hom} in this scenario, the latter explicitly adjusts for the trait correlations, which provides a well-behaved statistic under the null. The same explanation applies to **Figure 11b**, although here the loss in power for S_{Hom} is less pronounced due to the small genetic effect on the second trait.

For four or more traits, we use the 10 genetic effect vectors defined in **Table 2** of **Chapter 2**. **Figure 12** illustrates the results corresponding to $\boldsymbol{v_1}$, $\boldsymbol{v_4}$, $\boldsymbol{v_8}$ and $\boldsymbol{v_{10}}$ in relation to four phenotypes.

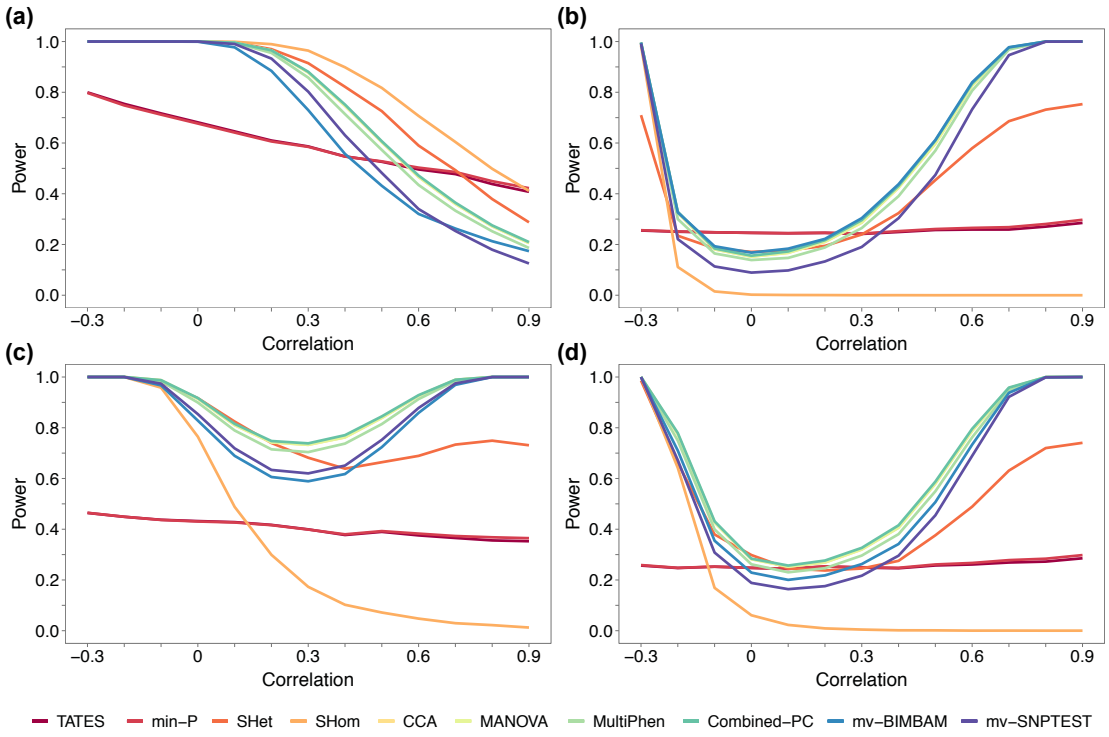

Figure 12. Power comparisons from simulations of scenario S1, based on **(a)** $\boldsymbol{v_1}$, **(b)** $\boldsymbol{v_4}$, **(c)** $\boldsymbol{v_8}$ and **(d)** $\boldsymbol{v_{10}}$ (see **Table 2** of **Chapter 2**) applied to data on four phenotypes. For all scenario S1 results the correlations between all phenotypes are the same. Correlations < -0.3 are not possible across four phenotypes, hence the truncation in these – and subsequent - results across the correlation range. Full results for scenario S1 are shown in **Figures 13 – 16**.

Figures 13 – 16 show the results for the remaining genetic effect vectors for 4 phenotypes and for all 10 genetic effect vectors in relation to 8, 20 and 48 phenotypes.

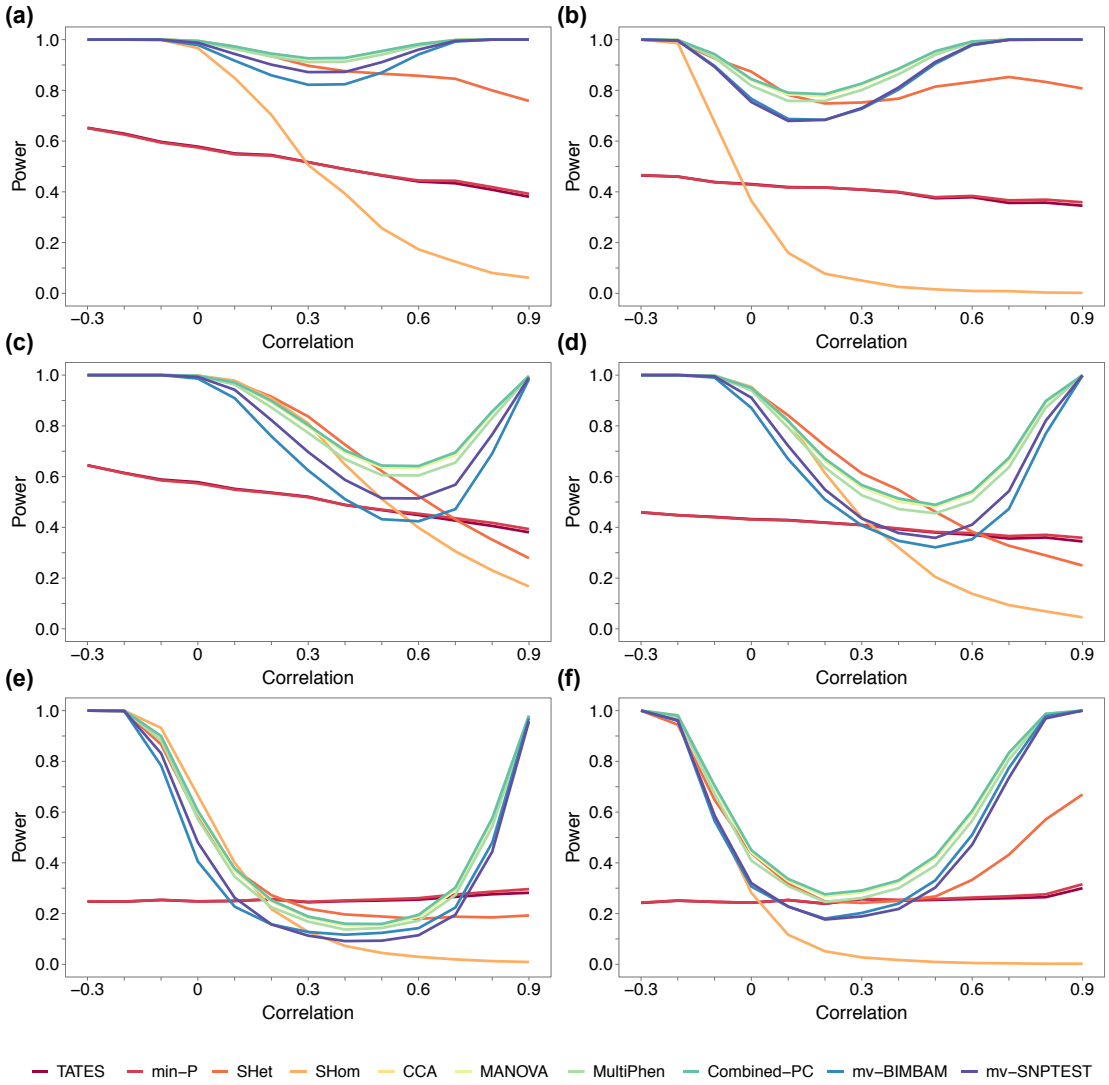

Figure 13. Power comparisons from simulations of scenario S1, based on **(a)** v_2, **(b)** v_3, **(c)** v_5, **(d)** v_6, **(e)** v_7 and **(f)** v_9 (see **Table 2** of **Chapter 2**) applied to data on four phenotypes. The pairwise phenotypic correlations are the same for all phenotypes. Correlations < -0.3 are not possible across four phenotypes, hence the truncation in these results across the correlation range.

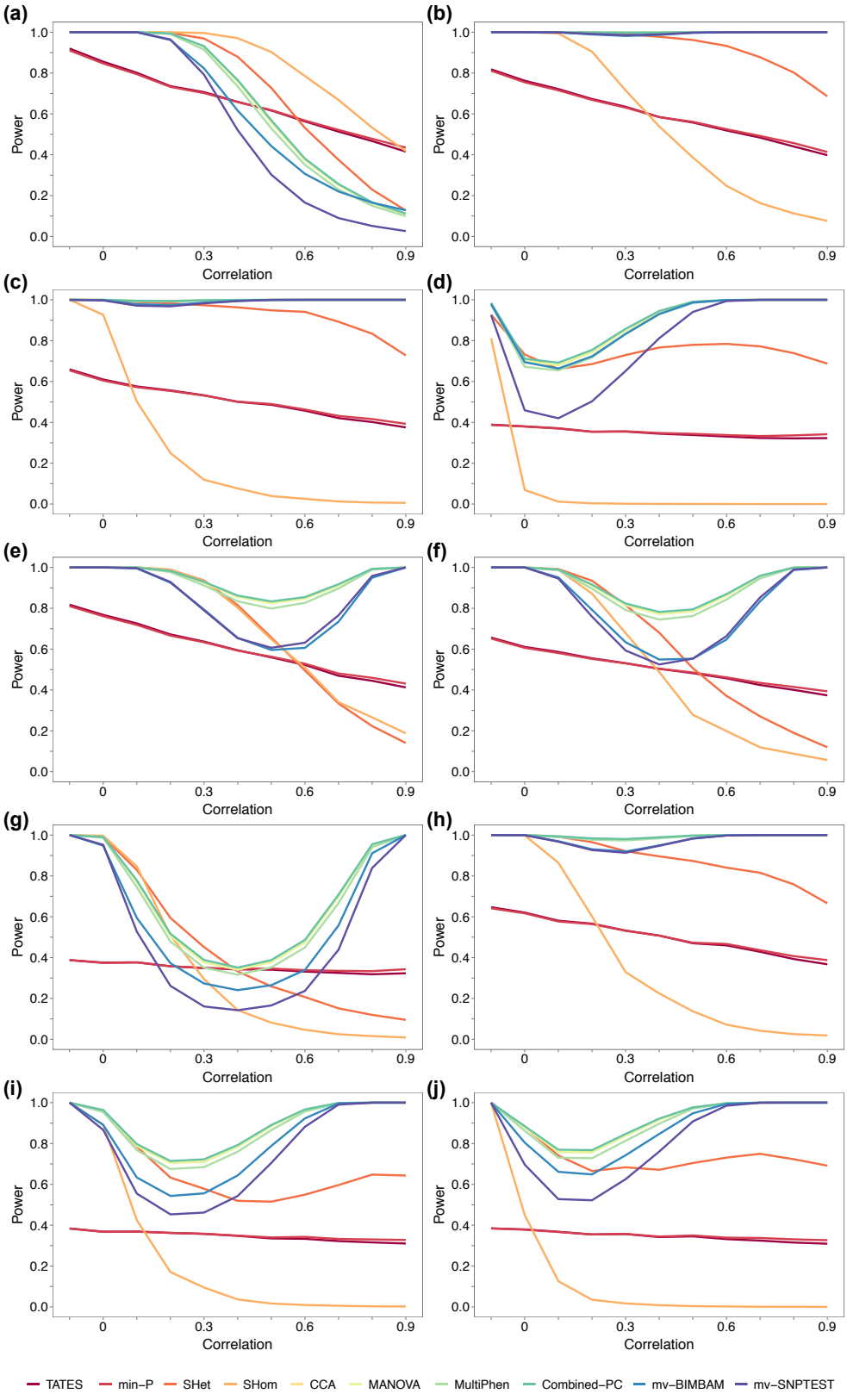

Figure 14. Power comparisons from simulations of scenario S1, based on $\boldsymbol{v_1}$ – $\boldsymbol{v_{10}}$ (see **Table 2** of **Chapter 2**) applied to data on eight phenotypes, **(a)** – **(j)** respectively. The pairwise phenotypic correlations are the same for all phenotypes. Correlations < -0.1 are not possible across eight phenotypes, hence the truncation in these results across the correlation range.

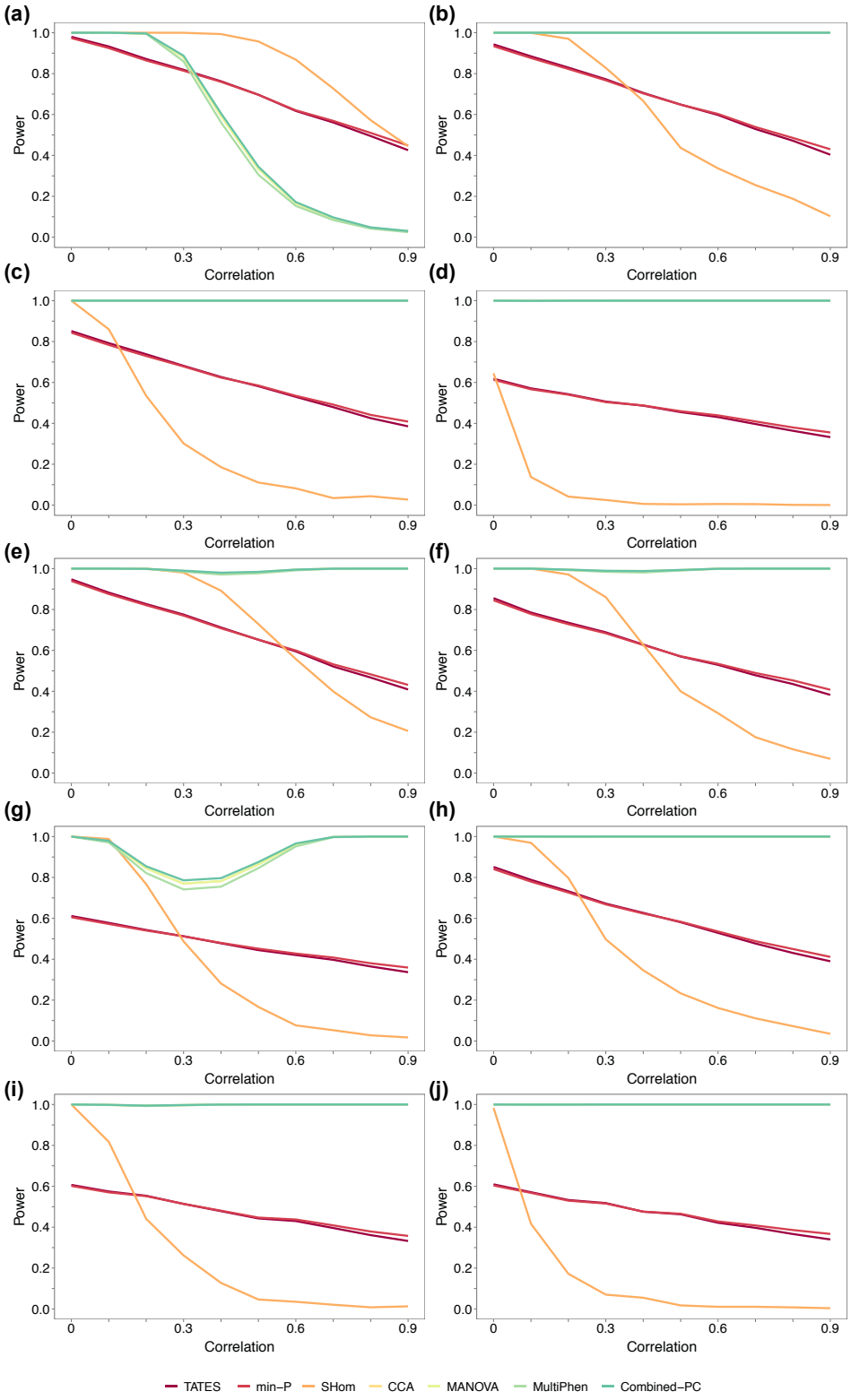

Figure 15. Power comparisons from simulations of scenario S1, based on $\boldsymbol{v_1}$ – $\boldsymbol{v_{10}}$ (see **Table 2** of **Chapter 2**) applied to data on 20 phenotypes, **(a)** – **(j)** respectively. The pairwise phenotypic correlations are the same for all phenotypes. Correlations < 0 are not possible across 20 phenotypes, hence the truncation in these results across the correlation range. mv-BIMBAM and mv-SNPTEST are not computationally feasible for 20 or more phenotypes and so are excluded here. S_{Het} is excluded, as a gamma distribution could not be estimated for these correlation matrices.

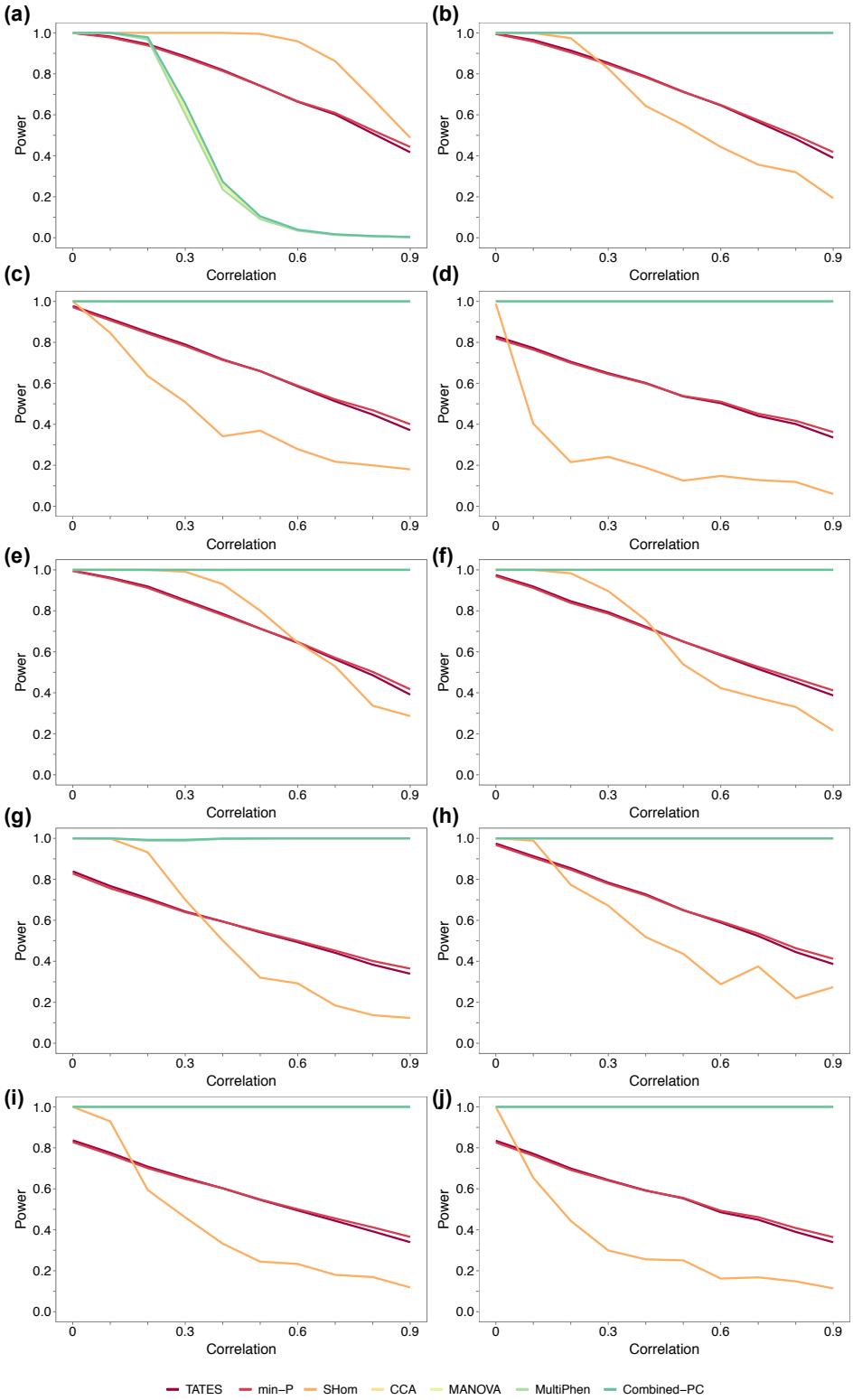

Figure 16. Power comparisons from simulations of scenario S1, based on $\boldsymbol{v_1} - \boldsymbol{v_{10}}$ (see **Table 2** of **Chapter 2**) applied to data on 48 phenotypes, **(a)** – **(j)** respectively. The pairwise phenotypic correlations are the same for all phenotypes. Correlations < 0 are not possible across 48 phenotypes, hence the truncation in these results across the correlation range. mv-BIMBAM and mv-SNPTEST are not computationally feasible for 20 or more phenotypes and so are excluded here. S_{Het} is excluded, as a gamma distribution could not be estimated for these correlation matrices.

A clear pattern emerges across these results. The individual-level methods form a 'leading group' in terms of power across much of the parameter space, in contrast to the lower performing pair of methods min-*P* and TATES, while S_{Hom} and S_{Het} tend towards this leading group the more pleiotropic the scenario (e.g. $\boldsymbol{v_1}$). S_{Het} has markedly higher power than the other summary statistic methods in most scenarios and often similar to the individual-level data methods. In the most pleiotropic scenario, $\boldsymbol{v_1}$, S_{Hom} performs best and min-*P* and TATES outperform the individual-level methods under high, positive phenotypic correlations (e.g. **Figure 11a**); otherwise S_{Hom} performs poorly. These differences in power between the methods increase with a greater number of phenotypes (see **Figures 12 - 16**).

Figure 17 shows the behaviour of the methods under the null hypothesis of no direct genetic effects on all phenotypes. While the methods generally perform as expected under the null, there is mild inflation for min-*P* and TATES under high phenotypic correlations for ≤ 8 phenotypes and for MultiPhen for 48 phenotypes, and strong deflation for min-*P*, TATES and S_{Hom} for 48 phenotypes. Therefore, use of these methods in these scenarios should either be avoided or else their statistics should be adjusted so that the error rates are controlled.

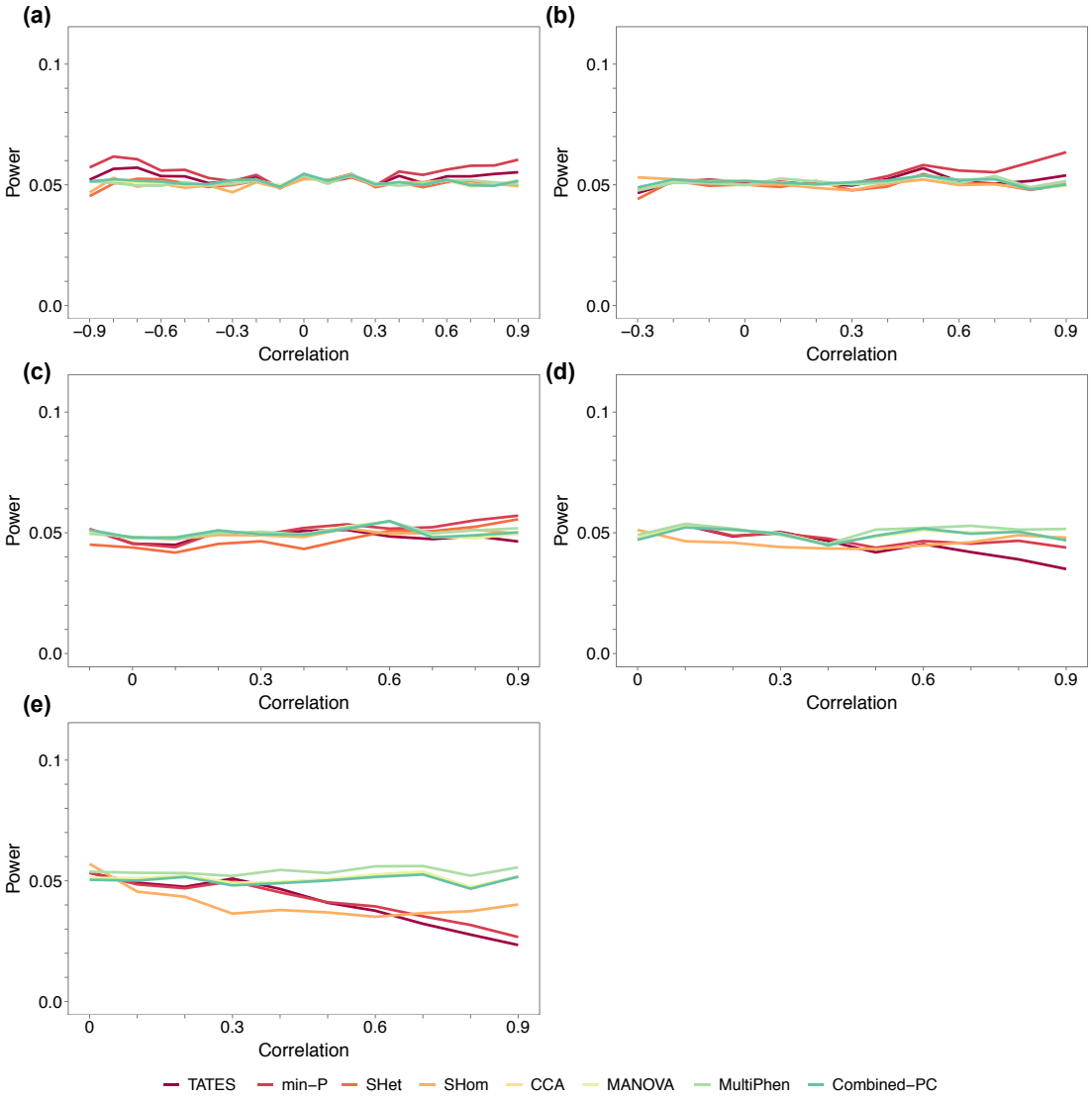

Figure 17. Simulations of scenario S1 under the null hypothesis of no genetic effect, applied to data on **(a)** 2, **(b)** 4, **(c)** 8, **(d)** 20 and **(e)** 48 phenotypes, based on 10,000 replicates. The pairwise phenotypic correlations are the same for all phenotypes, and the genetic variants are simulated to explain zero variance in all phenotypes. S_{Het} is excluded for 20 and 48 phenotypes, as a gamma distribution could not be estimated for these correlation matrices.

3.2.3.1 Downstream genetic effects

In addition, we perform simulations to test the performance of the multi-trait methods in the presence of indirect genetic effects, whereby the genetic variant has an effect on one of the tested phenotypes via its effect on another (a *downstream* or *mediated* effect). Here we perform simulations that model such an effect across two phenotypes (see **Chapter 2**). The results from these simulations (**Figure 18**) closely reflect those above in which there is a strong direct effect on one phenotype and no

or a small direct effect on the other. Given that a genetic effect on a trait will be sharply attenuated when it is only exerted via its effect on another tested trait, then unless the phenotypes are highly similar we expect our results on direct effects to capture the vast majority of those relating to indirect effects as well. One exception is the Combined-PC test, the results of which depart from the other individual-level data methods over some of the parameter space when the first trait has a very large effect on the second (see **Figure 18d**). Large differences in results between the Combined-PC test and the other individual-level data methods on real data may thus indicate the presence of an indirect effect, which may inspire a test to distinguish direct and indirect effects.

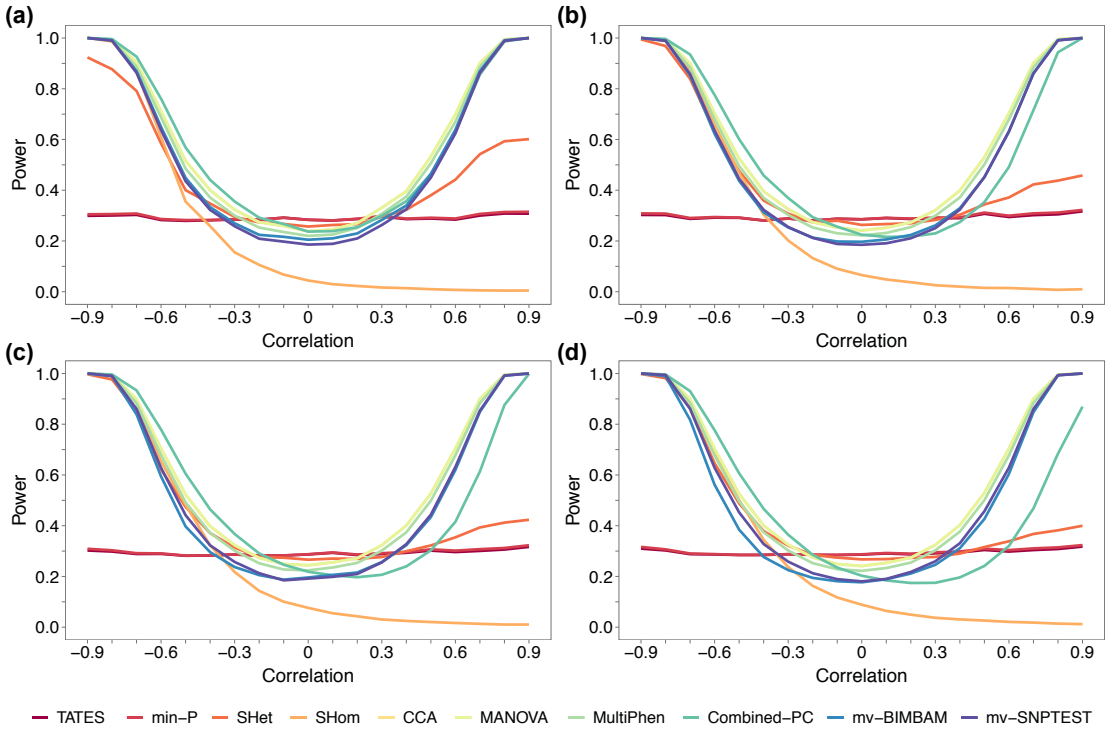

Figure 18. Power comparisons from simulations of scenario S1 applied to data on two phenotypes with simulated downstream genetic effects. Phenotypic variance explained by the genetic variant in trait 1 is 0.5% in all cases, and in trait 2 is **(a)** 1%, **(b)** 5%, **(c)** 10% and **(d)** 20%. The pairwise phenotypic correlations are the same for all phenotypes.

Figure 19 indicates that the behaviour of the methods under the null in the context of downstream effects reflects that for direct effects.

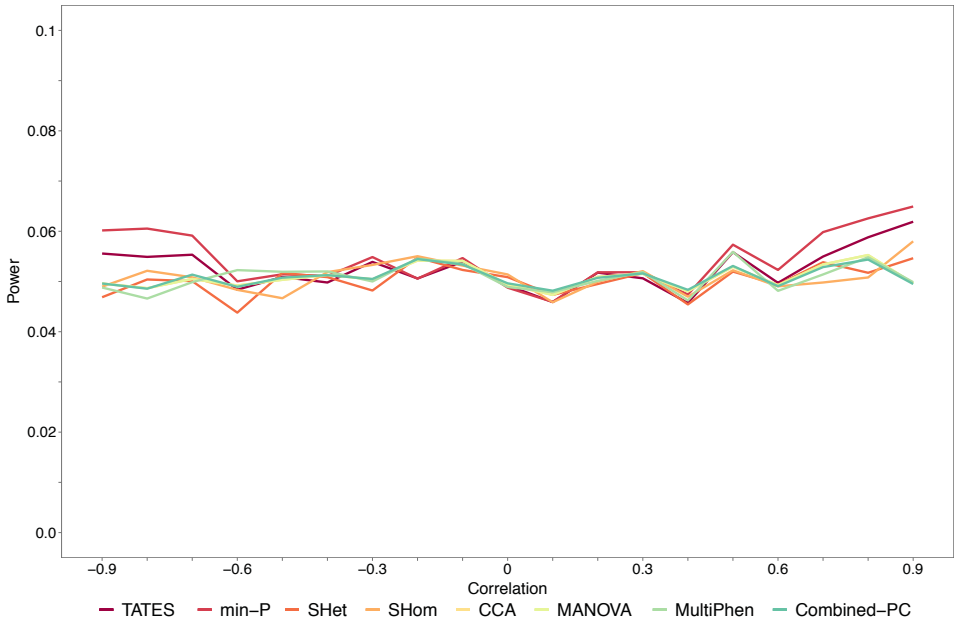

Figure 19. Simulations of scenario S1 with downstream effects under the null hypothesis of no genetic effect, applied to data on two phenotypes based on $10{,}000$ replicates. The pairwise phenotypic correlations are the same for all phenotypes, and the genetic variants are simulated to explain zero variance in the first phenotype, which has a downstream effect on the second phenotype.

3.3.1.2 Case-control phenotypes

Most of our simulation scenarios relate to quantitative traits, since the majority of the methods tested here are designed for the analysis of continuous variables. In scenario S1, we also simulate case/control phenotypes (see **Chapter 2**). **Figure 20** shows the results for the seven multivariate methods that can be applied to case/control data. The results show a similar pattern to those on quantitative traits only, with the individual-level data methods having similar power and greater than that of the summary statistic methods, apart from when the genetic effects reflect the phenotypic correlations. Min-*P* and TATES appear to perform relatively better when applied to case/control data in general but worse when the genetic variant affects both phenotypes equally, while S_{Hom} performs poorly when there is an effect on only one of the phenotypes, as expected.

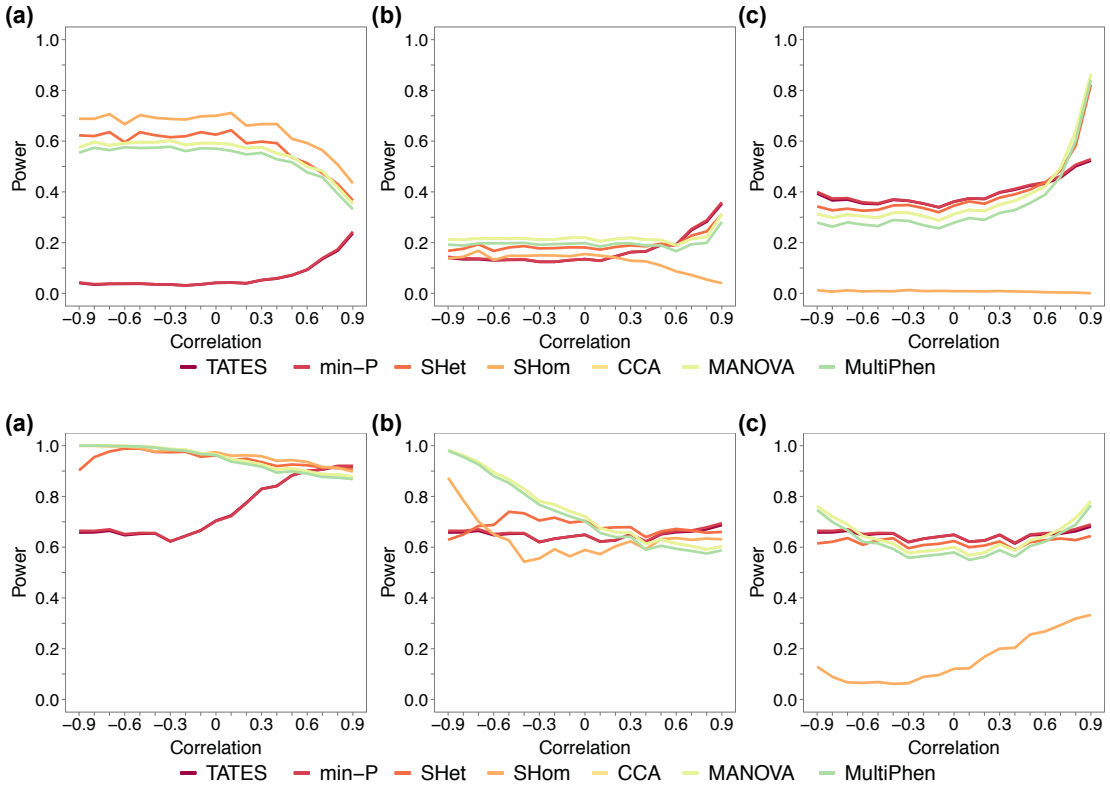

Figure 20. Power comparisons for the simulations of scenario S1 involving two case/control phenotypes (top panel), and one case/control phenotype and a quantitative phenotype (bottom panel). The genetic variant either has **(a)** the same effect on both phenotypes, **(b)** a larger effect on the first phenotype, or **(c)** an effect on the first phenotype and no effect on the second – in the mixed phenotype scenarios the first phenotype is the quantitative phenotype (see **Chapter 2** for details of these simulations). For all simulations, the case/control phenotypes have a simulated prevalence of 1% according to a liability threshold model.

3.3.1.3 Dissection of the Combined-PC method

The Combined-PC method (Aschard et al., 2014) performs a principal components analysis (PCA) on the phenotype data, and uses all PCs to test for association with the SNP. However, in other settings, such as controlling for population structure in GWAS, the top few PCs are often selected because they explain most variation in the data. This led us to investigate whether using just the top PC(s) may improve power. We performed simulations for two traits in scenario S1 and tested the Combined-PC method, as well as each PC separately. For the separate PC analyses, a linear regression was performed with the SNP as predictor and the PC as outcome. The same genetic effect vectors as in the earlier simulations of scenario S1 for two traits are used.

(a) Same genetic effect on both phenotypes

When the genetic variant affects both phenotypes equally and there is a negative phenotypic correlation, the second PC has highest power, mirroring the Combined-PC method (see **Figure 21**). For positive phenotypic correlations, we observe the opposite. Here PC1, which is in the direction of the positive phenotypic correlation, has optimal power to detect the genetic effect. The Combined-PC method performs as an average of both PCs here, in that it has high power for negative phenotypic correlations as a result of PC2, yet additionally exploits the power of PC1 for positive phenotypic correlations. However, this approach is worse than applying only the first PC for positive phenotypic correlations as power is lost by including PC2 in the model.

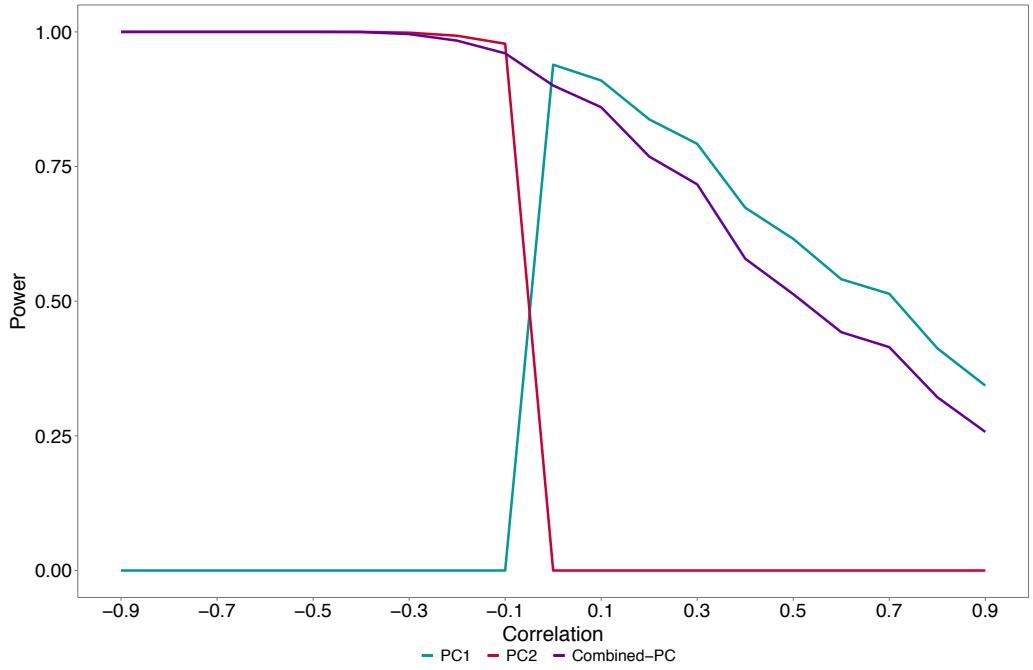

Figure 21. Power of the Combined-PC method, as well as the PCs individually under simulation of scenario S1 for two traits, where the genetic variant explains 0.5% variance in both traits.

Figure 22 illustrates pictorially why we observe this pattern of power of the two PCs in this scenario. When there is a negative phenotypic correlation (**Figure 22a**), PC1

is in that direction whereas PC2 is in the direction of the $Y = X$ line. Since there is a positive genetic correlation between the phenotypes in this scenario, the genetic effect is in the direction of the $Y = X$ line and is thus aligned to PC2, explaining its greater power for negative phenotypic correlations. When there is a positive phenotypic correlation (**Figure 22b**), PC1 is in the direction of the $Y = X$ line and is thus aligned to the direction of the genetic effect, resulting in greater power for PC1.

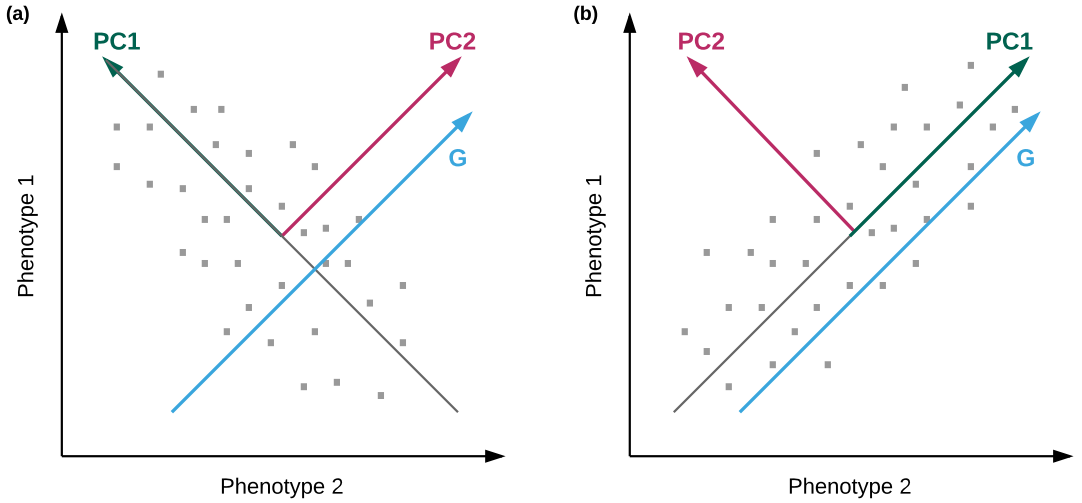

Figure 22. Illustration of the direction of the principal components (PC) and the genetic correlation (G) for two phenotypes where both phenotypes are affected by the genetic variant with the same magnitude.

(b) Different magnitudes of genetic effect

When the genetic variant affects both traits but with different magnitudes, again the first PC has minimal power to detect the causal association for negative phenotypic correlations (see **Figure 23**).

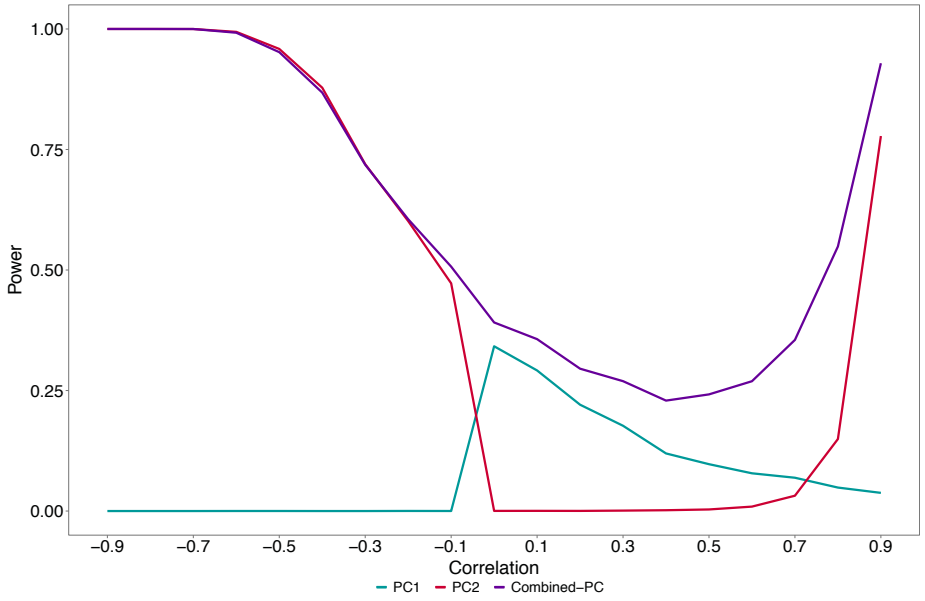

Figure 23. Power of the Combined-PC method, as well as the individual PCs under simulation of scenario S1 for two traits, where the genetic variant explains 0.5% variance in the first trait and 0.1% variance in the second trait.

For positive phenotypic correlations, however, PC1 largely has higher power than PC2. In this scenario a pictorial explanation for this pattern is complicated due to the different genetic effects on the phenotypes, and how the direction of the genetic effect compares to that of the PCs is challenging to explain intuitively. Thus, we consider the extreme scenario where only one trait is affected by the genetic variant (see below) to aid with the explanation of this scenario.

(c) Only one phenotype affected by the genetic variant

When the genetic variant affects only one of the two traits, the first PC has minimal power to detect the association between the SNP and the phenotype, whereas the power of the second PC follows a similar U-shaped pattern as the Combined-PC method, but overall has lower power (see **Figure 24**).

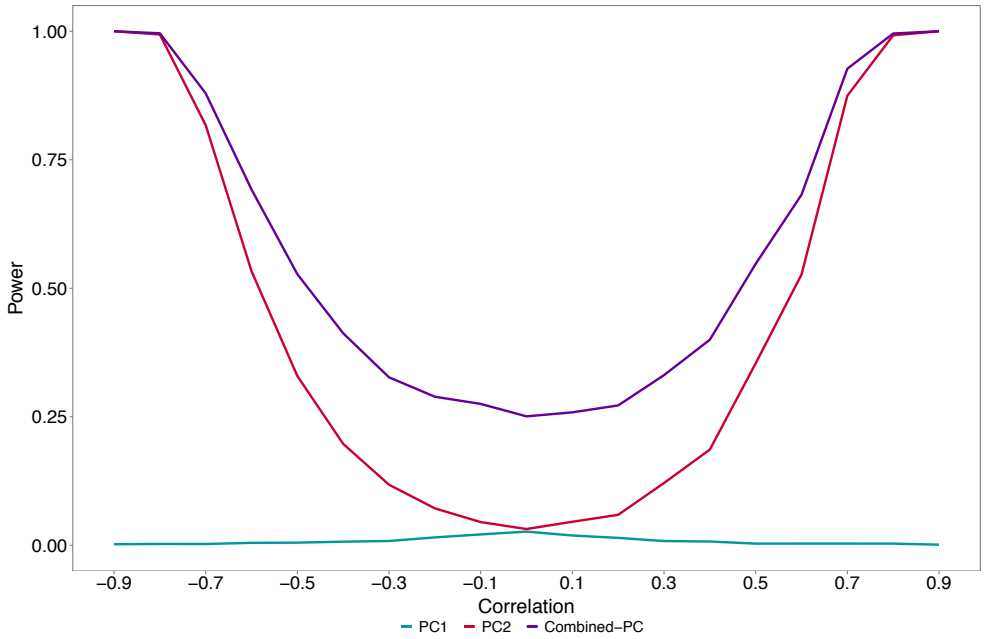

Figure 24. Power of the Combined-PC method, as well as the individual PCs under simulation of scenario S1 for two traits, where the genetic variant explains 0.5% variance in the first trait, and has no effect on the second trait.

Under this simulation setting, the genetic correlation between the traits is zero. If we imagine the extreme situation where the traits are perfectly correlated, such that their residual correlation is 1, the direction of this correlation would be completely aligned with PC1 on the line $Y = X$. However, one phenotype is affected by the genetic variant and there is no association between the genetic variant and the second phenotype. This additional genetic effect on the first phenotype leads to deviation from the perfect residual correlation, thus inducing spread around the $Y = X$ line, and so the genetic effect must be acting in the direction of the line $Y = -X$, which is the direction of PC2. Therefore, there is almost no power for PC1 and high power for PC2 here, which in this scenario has power equivalent to the Combined-PC method as the second PC is optimal. When the phenotypic correlation is weaker, this in itself induces spread around the $Y = X$ line, and so the power of PC2 decreases because the partial contribution of the genetic variant to the residual variance ensures that it is no longer perfectly aligned to PC2.

When there is a perfect negative correlation between the traits, such that their residual correlation is -1, the direction of the phenotypic correlation is aligned with the $Y = -X$ line, and so for the same reason as above, the genetic effect is aligned with $Y = X$, which is the direction of PC2 here. Given that when there are weaker levels of correlation between the phenotypes the direction of genetic effect does not align perfectly as in these extremes, the Combined-PC approach has optimum power by taking both PCs into account.

The scenario where there are effects on both phenotypes of different magnitudes, see section *(b)* above, is a less extreme version of the one considered here, where instead there is a small genetic effect on the second trait. In this situation we would observe that the genetic variant causes further deviation by its effect on the second phenotype, thus making the relationship between the PCs and the genetic variant even more complicated. Broadly we see that the power curves of **Figure 23** are similar to those of **Figure 21**, where the genetic variant has the same magnitude of effect on both traits, but in **Figure 23** we observe reduced power due to the diminished genetic effect on the second trait.

This investigation illustrates that the power of each PC is dependent on the genetic effects and phenotypic correlations, with one often being distinctly more powerful in a specific scenario. However, since the genetic effects are usually unknown, it is generally not possible to pre-select the most appropriate PC. As a general guide, when the genetic correlation aligns with the phenotypic correlation (both positive or both negative), the first PC has higher power. When the genetic and phenotypic correlations are not concordant, the second PC is preferable. However, when there exists weak genetic correlation, as in **Figure 23**, neither PC optimises power and the Combined-PC method is optimal. When more than two traits are analysed, the choice of the optimal PC becomes even more complex. Given that it is highly likely

that the traits to be analysed by such a method will contain a mixture of genetic effects of different magnitudes, the Combined-PC method will most likely, in general, optimise statistical power, as postulated in Aschard et al., 2014.

3.2.3.2 Investigating the effect of negative genetic correlations

In the previous simulations of scenario S1, while the magnitude of the simulated genetic effects was varied, the direction remained the same. Here we investigate the impact that a negative genetic correlation has on the power of the methods for simulations of scenario S1 with two traits, and for the genetic effect vectors defined earlier for this scenario. A negative genetic correlation in this scenario would be where the genetic variant increases the value of the first phenotype and decreases the value of the second phenotype (in the case of quantitative phenotypes), or vice versa. In terms of the data generating model as presented in **Equation 1** of **Chapter 2**, this corresponds to a positive beta β_1 for the first phenotype and a negative beta β_2 for the second, or vice versa. A reduced set of methods are used here as the findings for positive genetic correlations suggested two groups of method according to their performance – the univariate adjusted methods, here represented by min-*P*, and the individual-level data methods, here represented by CCA (mv-PLINK). In addition, S_{Hom} and S_{Het} expressed performance similar to the individual-level methods, but with performance differences in certain scenarios, hence why they are both included here. Further exploration of the effects of negative genetic correlations can be performed using our simulation tool.

(a) Same genetic effect on both phenotypes

The general effect on the power of the methods when there is negative genetic correlation is that the power curves of the positive genetic correlation simulations are reflected about a phenotypic correlation of zero (see **Figure 25**).

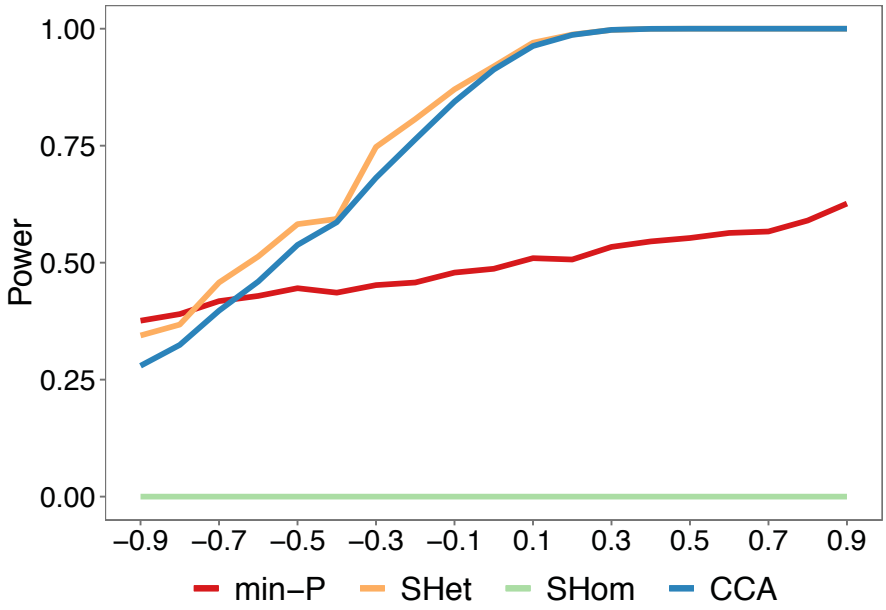

Figure 25. Power of the methods under simulation of scenario S1 for two traits, where the genetic variant explains 0.5% variance in both traits, but where the genetic effects are in opposite directions. The min-*P* method represents the performance of the univariate-adjusted methods; CCA (mv-PLINK) represents the performance of the individual-level methods.

However, the S_{Hom} method is an exception, producing zero power across the whole correlation range. As this method performs a meta-analysis of the traits, and the univariate t-values analysed are signed, the method has no power in the case of the same magnitude of genetic effects as these equal but opposite effects cancel each other out. This is not the case when the genetic effects have the same direction and thus the same direction of t-values.

(b) Different magnitudes of genetic effect

When the genetic correlation is negative - with genetic effects in opposite directions - and the magnitude of the genetic effects differ, the performance of all methods except S_{Hom} behave as expected (**Figure 26**), with the power curves a reflection of the ones observed above for positive genetic correlations. Unlike when the effects are equal and opposite, the S_{Hom} method does not have zero power across the whole

correlation range (see **Figure 26**), though overall the power is still substantially reduced.

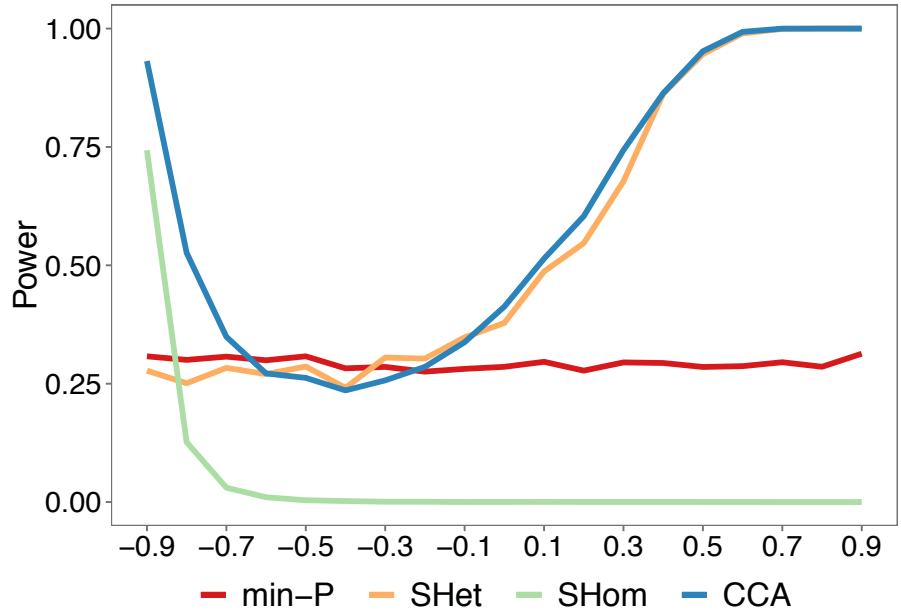

Figure 26. Power of the methods under simulation of scenario S1 for two traits, where the genetic variant explains 0.5% variance in the first trait and 0.1% variance in the second trait, but where the genetic effects are in opposite directions.

These observations provide an important consideration when performing multi-trait GWAS. It may be important to consider the direction of genetic effects before analysing multiple traits jointly, to avoid losing substantial power. In the application of GWAS summary statistics, we can observe the direction of effect of each SNP on each trait and could use this to guide which method to apply and which group of traits to analyse jointly.

We have shown that in some scenarios S_{Hom} has optimal power for detecting causal associations, but the results from the simulation of negative genetic correlations suggests that this method can also perform poorly. Therefore, we investigated ways in which to improve upon this loss of power by, for example, taking the absolute value of the t-values before applying the S_{Hom} method. The power of the methods

when absolute t-values are used for the S_{Hom} and S_{Het} methods is shown in **Figure 27** for two traits under scenario S1, as well as the min-*P* and CCA methods for comparison. **Figure 27a** shows the results for the simulation of the same magnitude of genetic effect and **Figure 27b** shows the results when the magnitude of the genetic effects are different.

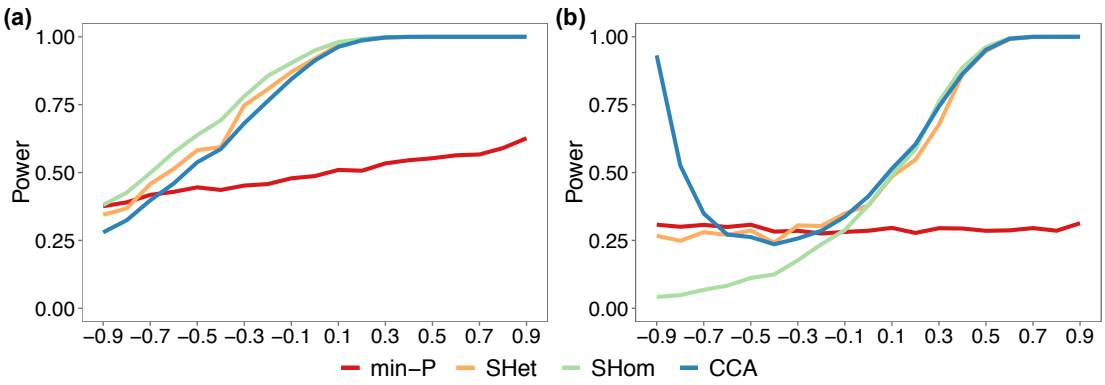

Figure 27. Power of the methods under the simulation of scenario S1 with two traits, where absolute t-values are analysed for the S_{Hom} and S_{Het} methods. **(a)** The genetic variant explains 0.5% variance in both traits. **(b)** The genetic variant explains 0.5% in one trait and 0.1% in the other trait. In both cases, the genetic effects are in opposite directions.

By taking the absolute of the t-values, we see that in both **Figure 27a** and **Figure 27b** the power of the S_{Hom} and S_{Het} methods are as expected; we observe the same pattern of power as in the case of positive genetic correlation, but here the curves are reflected about a phenotypic correlation of zero due to the opposite genetic and phenotypic correlations. However, when we perform null simulations by simulating no causal effect on both traits, this modification has led to the inflation of the S_{Hom} statistic, and in contrast, the deflation of the S_{Het} statistic (see **Figure 28**).

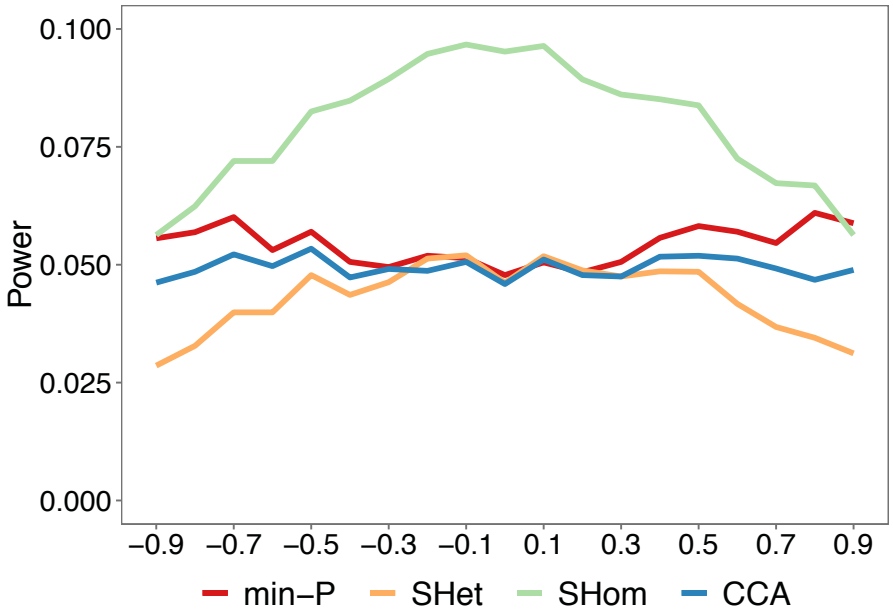

Figure 28. Power of the methods under the null simulation of scenario S1 with two traits, where absolute t-values are analysed for the S_{Hom} and S_{Het} methods.

Both the S_{Hom} and S_{Het} methods use the correlations between the t-values across traits as a proxy for the phenotypic correlations, and perform a t-value correlation and sample size weighted meta-analysis. Using the absolute t-values changes this t-value correlation, and since S_{Hom} follows a chi-squared distribution (with 1 degree of freedom), taking the absolute value results in its distribution departing from a chi-squared, leading to the observed inflation. S_{Het}, however, does not follow a standard distribution; instead a Gamma distribution is estimated via simulation.

The S_{Hom} method performs a test that is equivalent to one of the many tests performed in the S_{Het} procedure. S_{Het} tests subsets of the traits based on their univariate test statistics, including testing all traits given a specified threshold, whereas S_{Hom} only performs a meta-analysis on all traits. Taking the absolute t-values lead to deviation from the chi-squared distribution of the S_{Hom} statistic, thus we instead propose performing a version of the S_{Het} method where we specify the threshold such that all traits are analysed jointly for each SNP. Given that S_{Het} does

not appear to be affected by negative genetic correlations, we use the signed t-values in these simulations. The results of these simulations are provided in **Figure 29**.

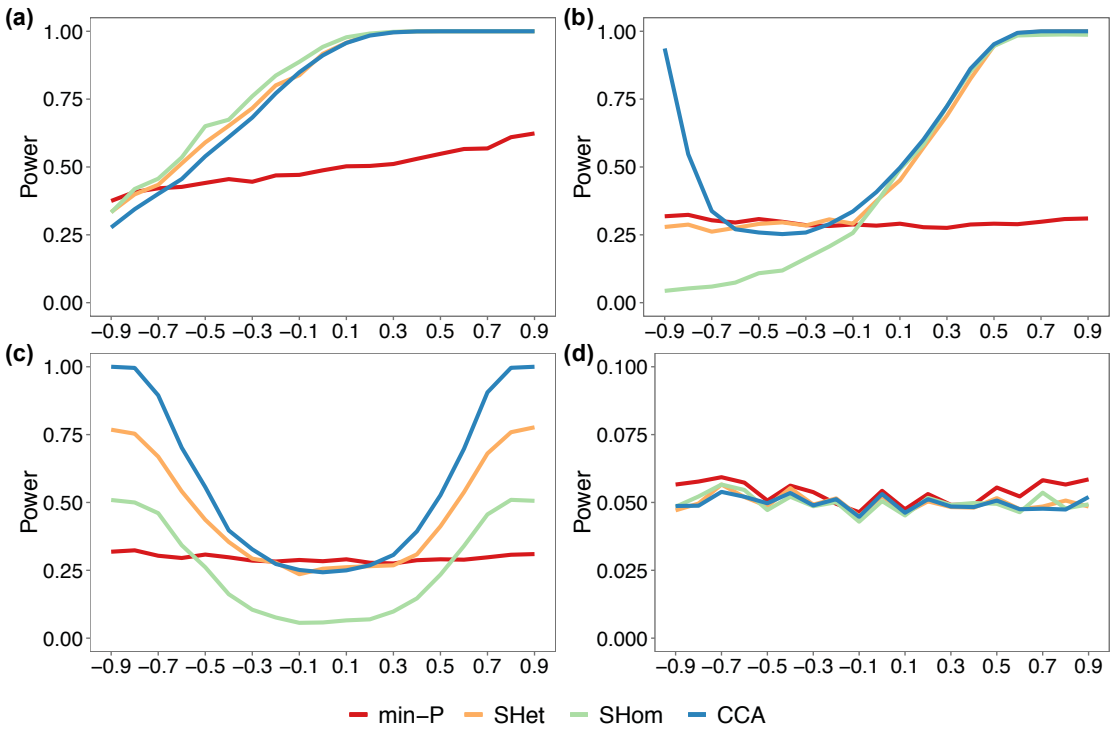

Figure 29. Power of the methods under the simulation of scenario S1 with two traits in which we replace the S_{Hom} method with a version of S_{Het} where we force all traits to be analysed jointly. **(a)** The genetic variant explains 0.5% variance in both traits. **(b)** The genetic variant explains 0.5% in one trait and 0.1% in the other trait. **(c)** The genetic variant explains 0.5% in one trait and has no effect on the other. **(d)** Null simulation where no genetic effects are simulated. In cases **(a)** and **(b)** the genetic effects are in opposite directions.

The power curves in **Figure 29a** and **Figure 29b** are as expected for negative genetic correlations and we see that S_{Hom} has equivalent power to when we simulate positive genetic correlations, reflected about a phenotypic correlation of zero as described above. We also performed simulations where the genetic variant affects only one of the traits (see **Figure 29c**). Here negative genetic correlations are not possible, but we are interested to find out how the new S_{Hom} approach performs for this combination of genetic effects. We observe that the power of this modified version of S_{Hom} follows the same pattern as the S_{Het} method, albeit with lower power;

this can be explained by the gain in power for the S_{Het} method by analysing subsets of traits when heterogeneous genetic effects exist.

The null simulation results under these settings are shown in **Figure 29d**, and we observe no inflation for the new S_{Hom} approach. These simulations suggest that, given that genetic effects are likely to be in different directions across the genome, this new S_{Hom} approach should be adopted in order to optimise the discovery power. In **Chapter 4** we utilise and develop this approach further, applying it to the latest GWAS summary statistics, but for the purpose of this comparison of published methods we revert back to applying the original S_{Hom} method for the remainder of the study.

3.2.4 S2: Genetic effects and phenotypic correlations sampled uniformly

In contrast to the structure of the S1 simulations, here we simulate data with genetic effects and phenotypic correlations sampled from uniform distributions, for 2, 4 and 8 phenotypes. **Figure 30** indicates that the individual-level data methods, as well as S_{Het} and S_{Hom}, have almost identical power and distinctly higher than that of min-*P* and TATES, with the difference larger for a greater number of traits. These results are in broad agreement with those of scenario S1, where this leading group of methods generally outperforms min-*P* and TATES.

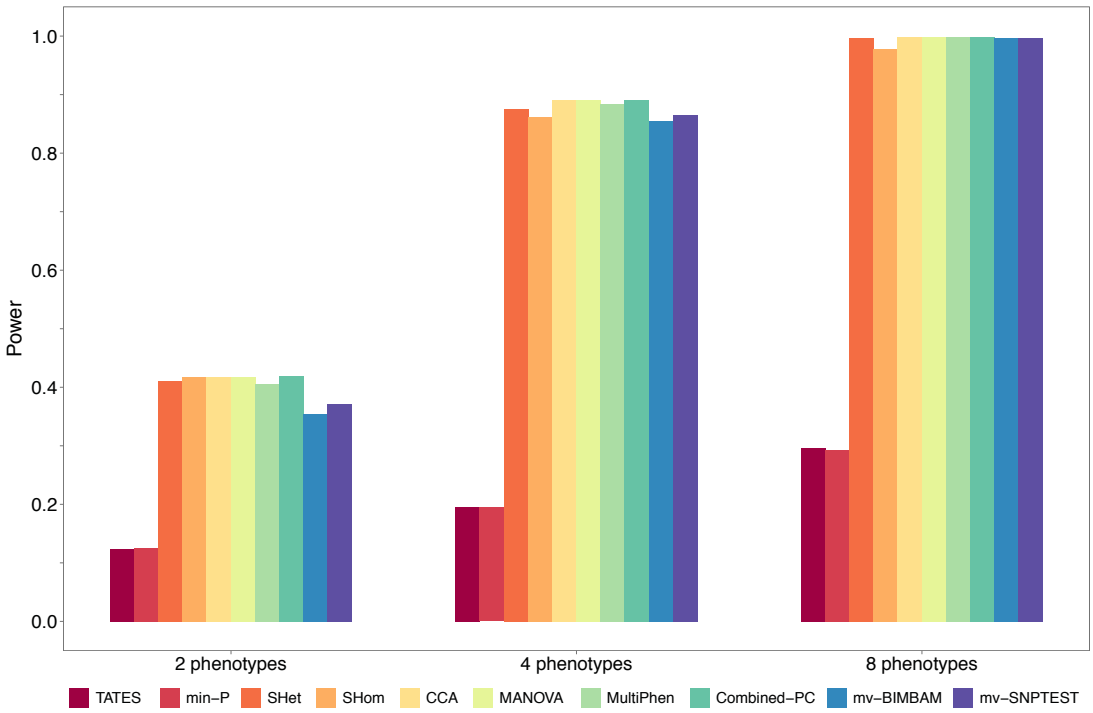

Figure 30. Power comparisons for the simulations of scenario S2 involving 2, 4 and 8 phenotypes. In this scenario the phenotypic correlations are chosen uniformly such that the correlation matrix is positive definite, and the effect sizes are sampled uniformly between 0% and 0.5% phenotypic variance explained.

3.2.5 S3: Genetic effects that reflect phenotypic correlations

In this simulation scenario we simulate genetic effects that are reflective of the phenotypic correlations, since it seems likely that on average the genetic correlations and phenotypic correlations are concordant.

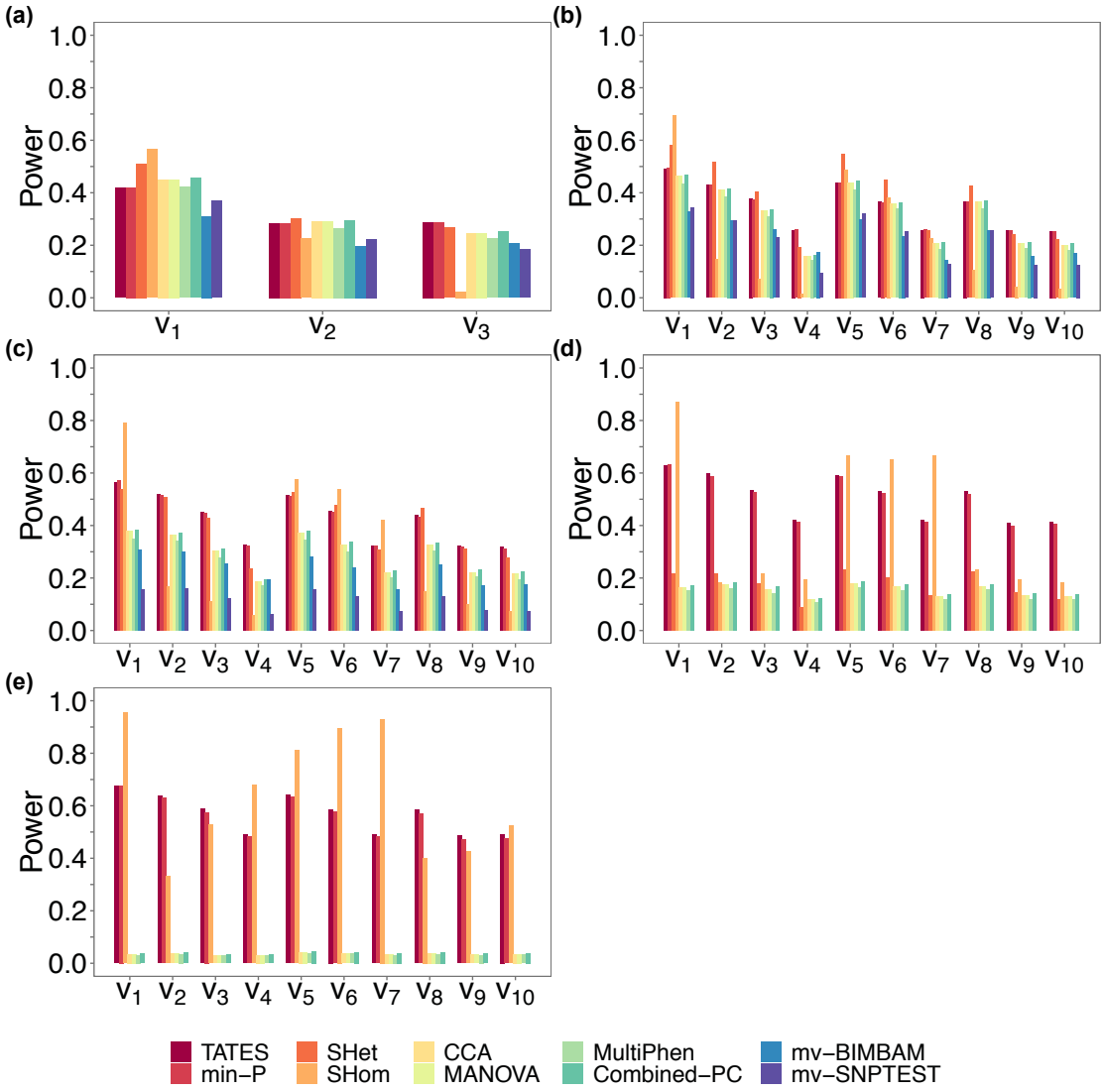

Figure 31. Power comparisons for the simulations of scenario S3 involving **(a)** 2, **(b)** 4, **(c)** 8, **(d)** 20 and **(e)** 48 phenotypes. In this scenario the phenotypic correlations are chosen to reflect the relative genetic effect sizes, as defined in **Chapter 2**. mv-BIMBAM and mv-SNPTEST are not computationally feasible for 20 or more phenotypes and so are excluded here for 20 and 48 phenotypes. S_{Het} is excluded for 48 phenotypes, as a gamma distribution could not be estimated for these correlation matrices.

While the results for two phenotypes are mostly similar across all methods (see **Figure 31a**), the summary statistic methods generally outperform the individual-level data methods more as the number of traits increases. The results of S_{Hom}, however, are sensitive to the genetic effect vector, being the best or worst performing summary statistic method depending on the vector, while the power of S_{Het} and the individual-level data methods is greatly reduced for 20 (see **Figure 31d**) and 48 (see

Figure 31e) traits. These results are in broad agreement with those of scenario S1 in which the genetic effects and phenotypic correlations are concordant.

3.2.6 S4: Real data informed simulations

This final simulation scenario derives the simulation parameters (genetic effects and phenotypic correlations) from published GWAS results, in order to generate data reflective of reality. This scenario is in two parts:

- **(a)** Real data informed phenotype correlations
- **(b)** Real data informed genetic effects and phenotype correlations

(a) Real data informed phenotype correlations

In this simulation scenario, we use the same fixed genetic effect vectors as in scenarios S1 and S3 (three for simulations of two phenotypes, and 10 for simulations of four or more phenotypes). Here, however, we sample pairwise phenotypic correlations from a multiple-Gaussian distribution that was fitted to a real correlation density based on NFBC1966 data (see **Chapter 2**).

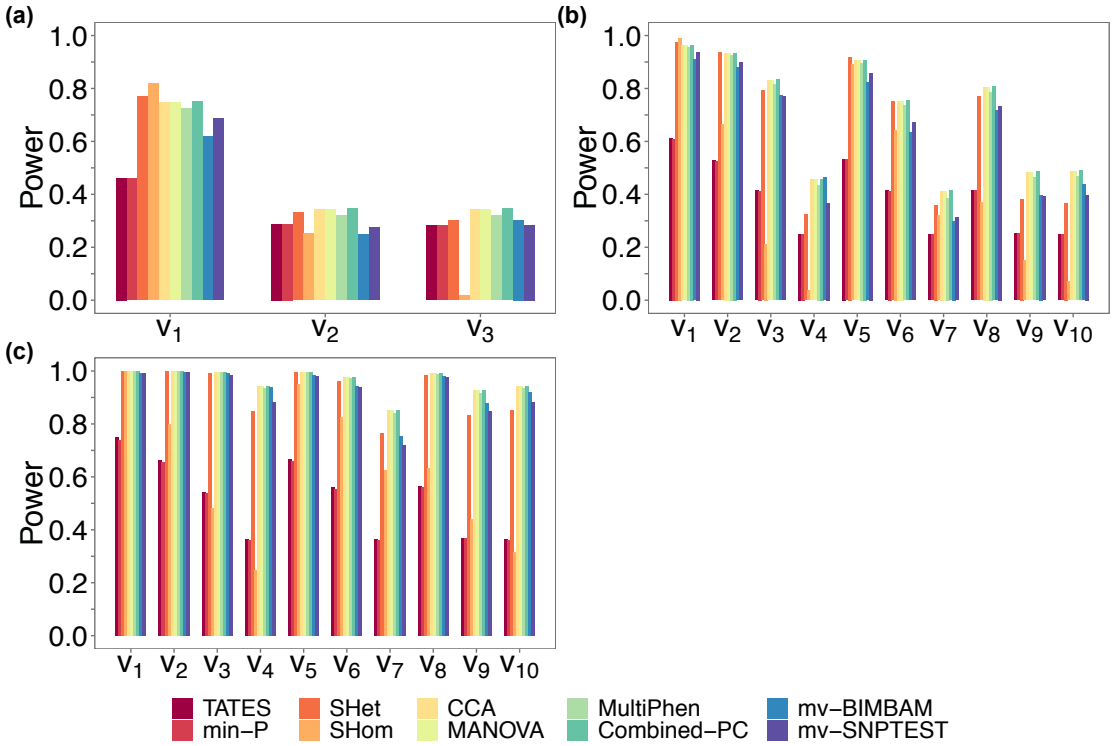

Figure 32. Power comparisons for the simulations of scenario S4a involving **(a)** 2, **(b)** 4 and **(c)** 8 phenotypes. In this scenario the phenotypic correlations are sampled from a fitted mixture Gaussian density (see **Chapter 2**), and genetic effect sizes are defined in **Table 2** of **Chapter 2**.

Figure 32 reveals that the individual-level methods and S_{Het} have markedly higher power than min-*P* and TATES for the majority of the genetic effect vectors. The performance of S_{Hom} is, again, highly dependent on the genetic effect vector, with greatest performance under pleiotropic effects. The results from this scenario are similar to those of scenario S2; in both cases the genetic effect vectors are independent of the phenotypic correlations, which optimises the statistical power of the individual-level data methods and S_{Het}.

(b) Real data informed genetic effects and phenotype correlations

In this simulation scenario we sample the genetic effects from real data, by utilising genetic effect estimates from published univariate GWAS (see **Chapter 2**). We obtain the pairwise phenotypic correlations directly from the NFBC1966 data on the corresponding traits.

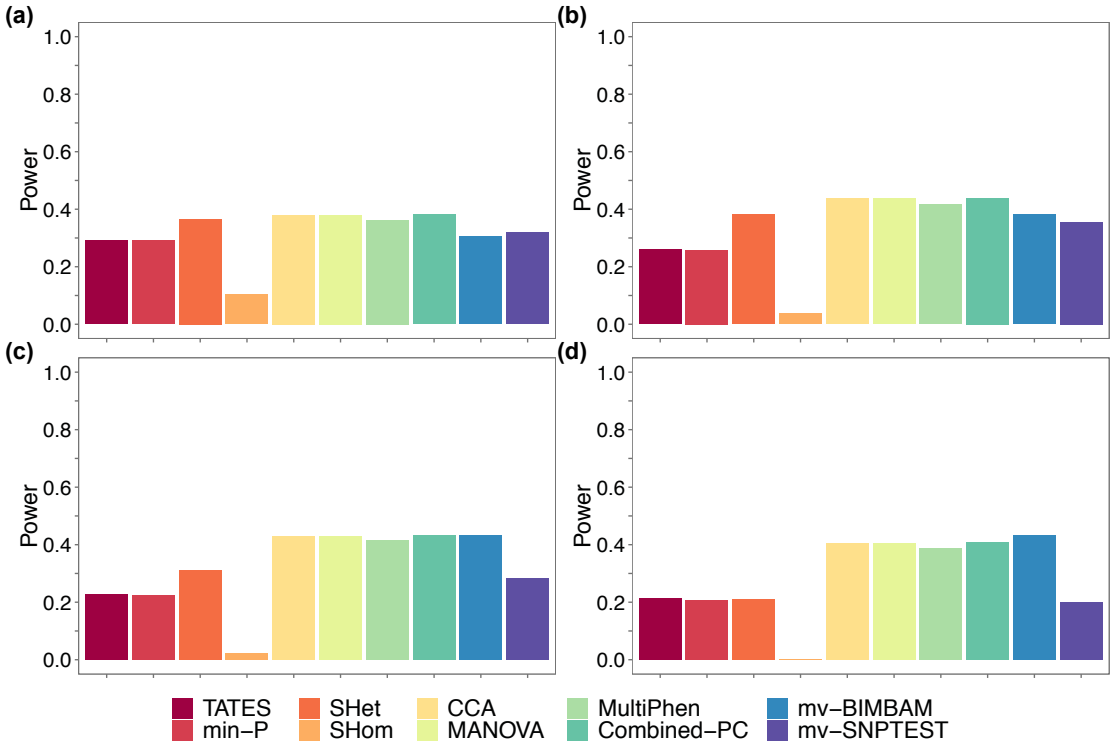

Figure 33. Power comparisons for the real data informed simulations of scenario S4b involving **(a)** 2, **(b)** 4, **(c)** 8 and **(d)** 12 phenotypes. For 12 phenotypes, all traits are analysed jointly. For 2, 4 and 8 phenotypes, data is simulated for all combinations of K phenotypes using the corresponding genetic effects and phenotypic correlations drawn directly from real data; the power results shown correspond to the average of the power estimates from all combinations.

These results, shown in **Figure 33** for 2, 4, 8 and 12 traits, may provide the most informative overall comparison of method performance given their basis on real combinations of effects and correlations. For two phenotypes, the individual-level methods substantially outperform min-*P*, TATES and S_{Hom}, while S_{Het} has similar power (see **Figure 33a**). As the number of phenotypes increases the power of the summary statistic methods decreases, with S_{Het} and S_{Hom} having the most dramatic decreases in power. For 12 phenotypes, S_{Het} has similar power to min-*P* and TATES, while S_{Hom} performs particularly badly in this scenario. From the results on 12 phenotypes we would expect the individual-level data methods to yield approximately twice the discovery of genetic variants than the summary data methods when applied to real data studies of the same sample size. However, for studies that can utilise much larger resources of summary statistic data than individual-level data, applying a

summary statistic method may optimise statistical power. To explore this, we simulated genotype-phenotype data relating to $10{,}000$ individuals and performed simulations using the best performing summary statistic method, S_{Het}, to evaluate the potential power gains. The results, shown in **Table 4**, indicate that S_{Het} has substantially higher power than the individual-level data methods at this increased sample size, although its advantage reduces with more traits. The expected power of the methods in studies exploiting individual-level or summary data of different sizes can be further estimated by our web application and software program that implement our simulation framework and the array of simulation scenarios described here.

Number of phenotypes	Power for 5,000 samples	Power for 10,000 samples	Maximum individual-level method power	Relative increase over individual-level method
2	0.366	0.947	0.382	148%
4	0.382	0.94	0.439	114%
8	0.31	0.894	0.432	107%
12	0.21	0.835	0.432	93%

Table 4. Power estimates for the S_{Het} method under simulation scenario S4b with simulated data on $5{,}000$ and $10{,}000$ samples. The maximum power achieved by any individual-level data method for $5{,}000$ samples is shown, as well as the percentage increase in power for the S_{Het} method on $10{,}000$ samples compared to this individual-level data method on $5{,}000$.

We also performed computation time calculations for each of the 10 methods in order to assess and compare feasibility of applying these methods. The results can be seen in **Table 5**, with computation time estimates given for 2, 4, 8 and 12 traits.

Method	2 traits	4 traits	8 traits	12 traits
min-*P*	0.001	0.001	0.001	0.001
TATES	0.072	0.118	0.228	0.438
S_{Het}	4.612	8.127	15.437	26.497
S_{Hom}	0.006	0.008	0.007	0.008
CCA	0.967	1.211	1.534	2.156
MANOVA	1.257	1.468	2.127	2.979
MultiPhen	49.036	55.135	72.09	108.078
Combined-PC	3.421	5.922	11.567	20.912
mv-BIMBAM	6.72	16.763	1186.968	57050.64
mv-SNPTEST	25.779	39.421	68.478	120.066

Table 5. Computation time estimates (in seconds) for the 10 methods for 2, 4, 8 and 12 phenotypes. We assessed the computation time for all 10 methods on a machine with a 2.7 GHz Intel Core i5 processor and 8 GB 1600 MHz DDR3 RAM. We simulated data for 5,000 samples, 100 SNP replicates with MAF 0.3, genetic variance explained of 0.5% for all phenotypes, and pairwise phenotypic correlations of 0.

As expected, the univariate adjusted method of min-*P* has the smallest computation time, as this method involves taking the smallest *P*-value and applying an adjustment for multiple testing, rather than directly performing a joint analysis. The computation time also does not increase much with the number of traits, making it scalable to large sets of phenotypes. The more complicated procedure of TATES does, however, incur additional computational burden with more phenotypes, and is consistently slower than min-*P*. Given the very similar performance between these two methods, min-*P* should be the preferred method for ease of implementation, as well as for the computational efficiency and scalability. Of the methods that utilise GWAS effect size estimates, S_{Het} is significantly slower than S_{Hom} due to the computationally intensive sub-setting procedure of the S_{Het} method. Since these methods have different optimal uses, one should not select one over the other. However, it could be the case that applying a manual sub-setting procedure, by choosing phenotypes to analyse jointly that are likely to exhibit homogenous genetic effects, and then applying the S_{Hom} method multiple times, is quicker than direct application of the S_{Het} method. Additionally, the computational load of S_{Het}

increases with more phenotypes, suggesting that the procedure described above could be a more feasible alternative for larger numbers of traits. While the individual-level method of mv-BIMBAM provides the greatest overall power in the real data simulations of scenario S4b for 12 phenotypes, and is computationally feasible for small numbers of traits, it has the largest computational time by more than two orders of magnitude in the case of just 12 phenotypes. The most computationally efficient individual-level method is CCA, and given the scalability of this method to more phenotypes and the similar performance to the other individual-level methods, CCA could be considered the individual-level method of choice given its ease of implementation. Computational feasibility considerations will become increasingly important when applying these methods to large multivariate panels, such as the UK Biobank. These computation time calculations were performed on 5,000 samples of 100 SNP replicates; for large panels of SNP data in population-wide cohorts, the computation time of these methods will be further compounded, making the choice of method based on computation time an important consideration. In the case of the biobank model, given we have observed that the summary statistic methods can replicate the power of the more computationally intensive individual-level methods in certain scenarios, it may be worth considering whether, for minimal loss in power, it would be more computationally efficient to perform univariate GWAS on the traits of interest before applying one of the summary statistic methods. Furthermore, by first obtaining the univariate summary statistic data, genetic correlation analyses can be performed in order to inform future joint analyses by establishing traits that are likely to exhibit pleiotropic effects, as well as enabling additional analyses such as polygenic risk scores.

Table 6 provides a summary of the performance, both statistical and computational, of the multi-trait GWAS methods in this comparison study.

Method		Performs well when…	Performs badly when…	Computational speed
Univariate	min-*P*	Genetic effects and phenotypic correlations are concordant	Genetic effects and phenotypic correlations are discordant	Fast and scalable to large numbers of traits
	TATES	Genetic effects and phenotypic correlations are concordant	Genetic effects and phenotypic correlations are discordant	Slow compared to min-*P* for similar power
Summary statistic	SHom	Pleiotropic genetic effects exist	Non-pleiotropic effects exist inc. different magnitudes and signs	Fast and scalable to large numbers of traits
	SHet	Genetic variant affects a subset of the traits under study	Analysing large numbers of traits due to multiple-testing penalty from sub-setting	Slow compared to some individual-level methods
Individual-level genotype-phenotype data	MANOVA	Genetic effects and phenotypic correlations are discordant	Genetic effects and phenotypic correlations are concordant	Fast and scalable to large numbers of traits
	CCA (mv-PLINK)	Genetic effects and phenotypic correlations are discordant	Genetic effects and phenotypic correlations are concordant	Fast and scalable to large numbers of traits
	MultiPhen	Genetic effects and phenotypic correlations are discordant	Genetic effects and phenotypic correlations are concordant	Comparatively slow for small numbers of traits
	Combined-PC	Genetic effects and phenotypic correlations are discordant	Genetic effects and phenotypic correlations are concordant	Relatively fast, but less scalable to more traits
	mv-BIMBAM	Genetic effects and phenotypic correlations are discordant	Genetic effects and phenotypic correlations are concordant	Moderate for few traits, but very intensive for many traits
	mv-SNPTEST	Genetic effects and phenotypic correlations are discordant	Concordant genetic effects and phenotypic correlations, and larger numbers of traits	Comparatively slow for small numbers of traits

Table 6. Summary of the performance and computational speed of the multi-trait GWAS methods included in the comparison study.

3.3 Discussion

Here we exploited the multivariate simulation framework presented in **Chapter 2** to perform a comprehensive comparison of the leading multi-trait GWAS methods. Development of multi-trait GWAS methodology has been an active area of research in recent years (Bottolo et al., 2013; Zhou and Stephens, 2012; O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Ferreira and Purcell, 2009; Klei et al., 2008; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007; Bolormaa et al., 2014; Casale et al., 2015; Huang et al., 2011; Zhang et al., 2014; Kim et al., 2016). However, publications introducing new methods are highly inconsistent in their evaluation of method performance, thus obscuring their relative merit. Across the range of simulation scenarios implemented, we were able to sufficiently expose the similarities and differences between the varying approaches, and provide insight into their performance. As well as acting as a guide to researchers performing multi-trait analyses, our findings will help to guide the future development of multi-trait GWAS methodology by highlighting areas where they are power gains to be made.

In the structured simulations of scenario S1, the individual-level data methods and the meta-analysis approaches of S_{Het} and S_{Hom} mostly outperform the univariate approaches of min-*P* and TATES. However, such a structured search of the model space may lead to testing unrealistic data, such as a genetic variant affecting only 12 of 48 traits, whose pairwise correlations are all 0.9. Therefore, some observed power differences may apply only to particular groups of traits or in settings outside genetic epidemiology. We also observed that when the genetic effects on the traits are in opposite directions (a negative genetic correlation), the meta-analysis approach of S_{Hom} loses a considerable amount of power. This highlights the importance of investigating the performance of different approaches so that optimal power can be

achieved by implementing the most appropriate method. When genetic effects and phenotypic correlations are sampled from uniform distributions (S2), S_{Het}, S_{Hom} and the individual-level data methods show markedly higher power than TATES and min-*P*. This is consistent with a general tendency for the individual-level data methods to have greatest relative power when the genetic effects and phenotypic correlations are discordant. This is further supported by the results from scenario S3, where the genetic effects and phenotypic correlations reflect each other. Here the summary statistic methods tend to perform best, especially S_{Hom} in the scenarios that are most pleiotropic. In the final scenario (S4b), genetic effects and phenotypic correlations are based on real data, and in the results relating to 12 traits the individual-level methods provide twice the discovery of genetic variants over the summary statistic methods.

Overall our results suggest that for a given sample size, the individual-level methods tested here are likely to optimise the discovery of genotype-phenotype associations. However, it should be noted that a summary statistic method with the same underlying assumptions as an individual-level method could be developed in the future, and thus reduce the gap in power between the two types of method. The choice of which individual-level method to use depends, in part, on the computational feasibility for the number of traits being analysed (see **Table 5**). For example, mv-BIMBAM has highest power in scenario S4b on 12 traits, but becomes computationally infeasible for a large number of traits (≥ 10). Other individual-level methods, in particular CCA, are preferable for a larger number of traits in terms of computation time (**Table 5**). The mv-BIMBAM method also provides additional interpretation by assigning probabilities to the combinations of direct, indirect and no effect of a SNP on the traits analysed, which provides insights into the genetic aetiology underlying multiple traits. If summary statistics are available on a sample that is markedly larger in size than that of available individual-level data then it is

highly likely that applying S_{Het}, in particular, will yield greatest power (**Table 4**). S_{Hom} is the best choice if the objective is to identify genetic variants with highly pleiotropic effects across all phenotypes under study.

In addition to providing a comprehensive guide to method choice in multi-trait GWAS, the extensive array of scenarios considered here expose several issues relating to the methods, not established in previous publications: (i) despite the sophisticated adjustment of univariate *P*-values performed by TATES, its power is approximately equivalent to simply adjusting the minimum *P*-value of the univariate tests by the effective number of independent tests (min-*P*); (ii) while the Combined-PC method shows almost equivalent power to the other individual-level data methods throughout, it has a marked departure in power for indirect genetic effects on the tested traits - this could provide a simple method for distinguishing direct and indirect effects; (iii) in many scenarios, S_{Het} and S_{Hom} have similar power to the individual-level data methods, demonstrating the potential for summary statistics to provide as much information as individual-level multivariate data; (iv) when genetic effects in opposite directions exist, however, S_{Hom} loses a considerable amount of power; (v) most multi-trait methods are not optimised for identifying pleiotropic variants (Solovieff et al., 2013), despite common reference to pleiotropy in publications that apply them; and (vi) S_{Hom}, which is tailored to detect pleiotropic variants, performs poorly in the real data informed simulations relating to 12 traits, suggesting that tests for pleiotropic variants may not produce novel findings unless applied to phenotypes that have prior knowledge of shared genetic aetiology.

While we have assessed the performance of many of the leading multi-trait GWAS methods across a range of different scenarios, we acknowledge that there are many scenarios not considered here. For example, we did not consider the situation where only one of many traits is affected by the genetic variant; in this setting we may

expect min-*P* and TATES to perform relatively well, despite our overall findings that these methods appear to have sub-optimal power. However, such simulations are easily performed using our simulation software, which implements the scenarios considered here and allows flexibility in user choice over parameters such as the number of traits, genetic effects and phenotypic correlations in order to simulate new scenarios. The results of this highly non-pleiotropic scenario, where we simulate 48 traits with only one causal association, are presented in **Figure 34**.

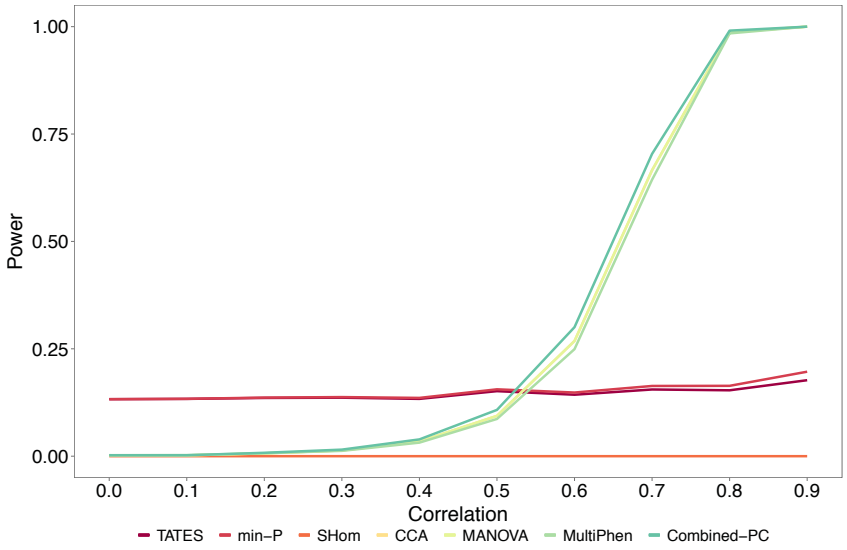

Figure 34. Power comparisons from simulations of scenario S1 for 48 traits, where only one trait is affected by the genetic variant. The genetic variant is simulated to explain 0.5% variance in the affected trait. The pairwise phenotypic correlations are the same for all phenotypes. Correlations < 0 are not possible across 48 phenotypes, hence the truncation in these results across the correlation range. mv-BIMBAM and mv-SNPTEST are not computationally feasible for 20 or more phenotypes and so are excluded here. S_{Het} is excluded, as a gamma distribution could not be estimated for these correlation matrices.

Min-*P* and TATES do indeed perform relatively well in this scenario, although the individual-level methods have greater power for higher phenotypic correlations. We also recognise that there are other multivariate approaches to genetic association studies that have not been considered here, such as linear mixed models (Zhou and Stephens, 2012), generalised estimating equations (Zhang et al., 2014) and adaptive testing (Kim et al., 2016), as well as multi-SNP, multi-trait approaches (Bottolo et al.,

2013; Zhou and Stephens, 2012; Kim et al., 2016). It is possible that under some assumptions these methods are preferred over those considered here, though our results suggest that the power of most multivariate methods converge to some optimal value for a large part of the model space. Additional methods can be easily incorporated into our simulation software, allowing further methods to be benchmarked under a wide range of simulation settings.

Multivariate genetic analyses are likely to expand dramatically in future as an increasing number of GWAS results are released publicly and as individual-level multivariate panels are compiled by population-wide biobank studies. This makes our study extremely timely, and designing studies guided by its findings should lead to greater discovery of genuine associations. This could be especially significant for underpowered but extremely important phenotypes, such as depression (Hyman, 2014; Ripke et al., 2013; Converge Consortium, 2015; Hyde et al., 2016), for which few genetic associations have been discovered. Multivariate methods may leverage the power of GWAS on such phenotypes, providing vital targets for drug development in diseases and disorders with few biological leads. In addition to the direct benefits of increasing the number of known genetic associations for any phenotype, without cost, this will also produce higher-powered downstream analyses, such as pathway analyses and polygenic risk scoring.

While our results provide a present snapshot of multi-trait GWAS method performance, our simulation framework offers a consistent platform from which future methods can be easily benchmarked via our web application and open-source software program. This should save researcher time and avoid repetition by guiding the development, application and publication of only those methods demonstrated as outperforming the alternatives. We believe that this study highlights the importance of

systematic and comprehensive comparisons of competing methods of analysis, easily reproduced and extended via open-source software.

4. Identifying novel loci from multi-trait GWAS on summary statistics

In **Chapter 3**, we observed that multi-trait summary statistic GWAS methods could achieve equivalent power to individual-level methods, and even greater power when taking into account that larger resources of summary data are now available. These findings provide the motivation for this chapter, where we perform a series of multi-trait GWAS using publicly available summary data on 19 quantitative and binary phenotypes. We develop and apply different approaches to multi-trait summary statistic GWAS, which aim to capture both homogenous and heterogeneous genetic effects on sets of correlated traits. While one aim of this study is to confirm the results of our simulations, showing that novel findings can be achieved by repurposing existing summary data in multi-trait analyses, we also aim to make new discoveries from real data that can provide insights for both the study of specific phenotypes and for the study of pleiotropic genetic loci.

4.1 Introduction

Genome-wide association studies (GWAS) have been performed across hundreds of phenotypes leading to the identification of thousands of SNPs associated with human phenotypes. These results and those from the application of methods for estimating genetic correlation between traits, such as bivariate GCTA, LD Score regression and others (Yang et al., 2011; B. Bulik-Sullivan et al., 2015; B. K. Bulik-Sullivan et al., 2015; Han et al., 2016; Pickrell et al., 2016) have demonstrated a substantial enrichment of variants affecting multiple traits (see: *A plethora of pleiotropy across complex traits* (Visscher and Yang, 2016)). Therefore, performing GWAS on phenotypes jointly could increase the discovery of susceptibility variants and provide insights into pleiotropic aetiology, yet GWAS to date have typically analysed phenotypes independently. However, the public release of large-scale GWAS summary statistics across a large and growing number of phenotypes motivates the development and application of multi-trait GWAS methods that exploit summary data.

Summary statistic methodology development is an emerging area of research in statistical genetics, with many studies now seeking to exploit univariate GWAS summary statistics to gain further insight into the genetic aetiology of complex disorders. An already highly popular method, known as LD Score regression (B. K. Bulik-Sullivan et al., 2015), has been used to produce an 'atlas' of genetic correlations among a range of phenotypes with publicly available GWAS data (B. Bulik-Sullivan et al., 2015). Another recent study investigated the shared genetic aetiology of 42 traits by performing a genome-wide scan to detect genetic variants that affect pairs of traits, assessing the causal relationship between these traits (Pickrell et al., 2016). Moreover, a review into the advancement of summary statistic methodology has recently been undertaken, highlighting that for many applications

summary statistic approaches may be preferred over the individual-level approach due to the gains in sample size and thus power (Pasaniuc and Price, 2016).

Here we develop highly-powered multi-trait GWAS methods that utilise summary statistics, and that are appropriate for a mixture of binary and continuous traits. We perform simulations to compare their relative performance against other summary statistic GWAS methods and to the individual-level multi-trait approach. We apply these methods to GWAS summary statistic data on a range of traits, including anthropometric, metabolic and psychiatric traits. We identify novel findings from our multi-trait analyses, which can provide further insights into pleiotropy and the biology underlying correlated traits.

4.2 Materials and Methods

4.2.1 Multi-trait GWAS methods

In order to compare the performance of the different summary statistic approaches, we perform simulations utilising the simulation framework presented in **Chapter 2**. We compare MetaHom, MetaHet, developed here, and the 'Cattle method' (Bolormaa et al., 2014), to min-*P* (O'Reilly et al., 2012) and TATES (van der Sluis et al., 2013), which both adjust univariate *P*-values to obtain a joint association *P*-value. The final two methods do not explicitly model the traits jointly as they only correct the univariate *P*-values, and by definition they cannot identify novel findings, but we include them here to compare to the univariate approach and to act as a benchmark for the joint multi-trait methods. Furthermore, we include the individual-level multi-trait method CCA (mv-PLINK) (Ferreira and Purcell, 2009), chosen here as a representation of individual-level method performance.

4.2.1.1 MetaHom

This test extends the S_{Hom} statistic introduced by Zhu *et al.* (Zhu et al., 2015), making it applicable to a mixture of binary and continuous phenotypes (see **4.2.2**).

The S_{Hom} test statistic is given by:

$$S_{Hom} = \frac{e^T(RW)^{-1}S(e^T(RW)^{-1}S)^T}{e^T(WRW)^{-1}e}$$

where S is the matrix of summary statistics, R is the correlation matrix between the summary statistics, W is the diagonal matrix of weights for each phenotype under study, where $w_{ij} = \sqrt{n_i}$ for sample size n_i of the i^{th} cohort, and e is the identity vector of length $i \times j$. S_{Hom} follows a chi-squared distribution with one degree of

freedom.

S_{Hom} combines Wald test statistics from univariate GWAS relating to a SNP across both multiple cohorts and multiple phenotypes in a meta-analysis. Heterogeneity in effect size and statistical power across cohorts is accounted for, as is the correlation among the test statistics, while the overall test statistic has optimal power when the genetic effect is homogeneous across traits and cohorts. Of the tests considered here, S_{Hom} is that which can be most considered a 'test for pleiotropy', being equivalent to a meta-analysis of effect sizes across traits and cohorts with optimal power under fixed effects.

As identified in **Chapter 3**, when the direction of genetic effects of a SNP on multiple traits are in opposite directions, the power of S_{Hom} is greatly reduced due to the effective 'cancelling-out' of the genetic effects in the meta-analysis. However, we do not consider it appropriate to make this fixed effects assumption in a multi-trait GWAS setting, because we seek to identify any SNP that has a non-null association on a group of traits irrespective of the direction of any effects on those traits. Thus, our MetaHom method is in fact a special case of MetaHet (see below), which does not make a fixed effects assumption but searches for SNPs with a similar *absolute* effect on the set of phenotypes.

4.2.1.2 MetaHet

This test extends the S_{Het} statistic introduced by Zhu *et al*. (Zhu et al., 2015) making it applicable to a mixture of binary and continuous phenotypes (see **4.2.2**). This test is derived from S_{Hom} but is designed to detect genetic variants that only affect a subset of the total number of traits under study. In addition, the signs of the weightings are adjusted to take into account the direction of genetic effect such that the effects add rather than cancel out as in the S_{Hom} meta-analysis.

For each subset of traits, the S_{Het} method performs a meta-analysis similar to that of S_{Hom}, but with signed weightings so that the absolute effect sizes are considered. For each subset, determined by the summary statistic threshold for inclusion τ, the corresponding statistic is given by:

$$S_{Het}(\tau) = \frac{e^T(R(\tau)W(\tau))^{-1}S(\tau)(e^T(R(\tau)W(\tau))^{-1}S(\tau))^T}{e^T(W(\tau)R(\tau)W(\tau))^{-1}e}$$

where $S(\tau)$ is the sub-matrix of summary statistics, $R(\tau)$ is the sub-matrix of the summary statistic correlation matrix R, $W(\tau)$ is the diagonal matrix of weights for the subset of phenotypes under study, where $w_{ij} = \sqrt{n_i} \times sign(S_{ij})$ for sample size n_i of the i^{th} cohort and summary statistic S_{ij} for the j^{th} phenotype in the i^{th} cohort, and e is the identity vector of length $i \times j$. For each summary statistic threshold τ the phenotypes satisfying $|S_{ij}| > \tau$ are selected for joint analysis.

The maximum value of $S_{Het}(\tau)$ across all analysed subsets is then taken to be the S_{Het} test statistic:

$$S_{Het} = \max_{\tau>0} S_{Het}(\tau)$$

As the S_{Het} test statistic does not follow a standard distribution, *P*-values are computed via simulation of an empirical gamma distribution.

In the MetaHet implementation of the S_{Het} method here, we use the default option for the sub-setting procedure: the traits are ordered according to the absolute value of their Wald test statistics, then for each SNP the association with subsets of the traits is tested recursively, starting with the trait with the largest absolute t-statistic, and progressively adding in one trait at a time until the SNP is tested for association with all traits. Both S_{Het} and S_{Hom} are unaffected by sample overlap among meta-analysed studies (Zhu et al., 2015).

The software implementation of the S_{Het} method allows for the sub-setting procedure to be bypassed, thus for all traits to be analysed jointly as in the S_{Hom} procedure. Simulation results presented in **Chapter 3** demonstrated that the S_{Hom} method loses power when applied in real data scenarios due to the cancellation of opposite genetic effects in the meta-analysis. We showed that implementing the S_{Het} method with the sub-setting procedure disabled replicates the power of the S_{Hom} method, and is unaffected by oppositely signed genetic effects due to the signed weightings. Hence, we propose this modification to the S_{Het} method to perform a test similar to that of S_{Hom} that does not lose power when jointly analysing traits with negative genetic correlations. The MetaHom method implemented here is a special case of MetaHet, where all traits are analysed jointly.

4.2.1.3 Cattle

The 'Cattle' summary statistic method (Bolormaa et al., 2014) performs a t-value correlation weighted meta-analysis, and has the following test statistic:

$$S_{Cattle} = S\,(cor(S))^{-1}\,S^{T}$$

where S is the matrix of t-values from univariate GWAS. This test statistic follows a chi-squared distribution with K degrees of freedom, where K is the number of traits. This method was compared to the S_{Het} and S_{Hom} methods, and was noted to perform poorly for heterogeneous genetic effects, whereby a SNP affects a subset of the traits under study, due to the large number of degrees of freedom of the test-statistic (Zhu et al., 2015). This method has thus far not been applied to human traits, with the publication presenting this method performing analyses of Cattle specific traits (Bolormaa et al., 2014).

4.2.1.4 Comparison methods

We compare the performance of the three summary statistic methods detailed above to three additional approaches to multi-trait GWAS, which were described in more detail in **Chapter 3**. Min-*P* (O'Reilly et al., 2012) performs an adjustment of the univariate *P*-values to obtain a joint multi-trait *P*-value. Since only the univariate *P*-values are adjusted, and the phenotypes aren't explicitly modelled jointly, by definition this method cannot produce novel findings. Similarly, the TATES method (van der Sluis et al., 2013) adjusts univariate *P*-values using the eigen-decomposition of the phenotype correlation matrix, but like min-*P* cannot identify novel findings. CCA (mv-PLINK) (Ferreira and Purcell, 2009) is an individual-level multi-trait GWAS method, that is equivalent to a reversed multiple linear regression with a single genetic variant as outcome (O'Reilly et al., 2012). The performance of the CCA method in the simulations represents the performance of the individual-level multi-trait GWAS methods.

4.2.2 Effective sample size for case/control studies

Both the MetaHet and MetaHom methods use, in addition to the correlation between the t-values, the sample sizes of the univariate studies to weight the t-values as a way of quantifying the certainty in the beta estimates. We use the following formula to convert sample sizes from case/control studies into quantitative trait equivalents:

$$N_{eff} := \frac{4}{\left(\frac{1}{N_{cases}} + \frac{1}{N_{controls}}\right)} \quad \textbf{(3)}$$

where N_{eff} is the effective quantitative sample size of a case/control study in terms of statistical power, N_{cases} is the number of cases and $N_{controls}$ is the number of

controls in a case/control study (Han and Eskin, 2011; Ma et al., 2013). For example, if a GWAS were performed for a binary trait on $10{,}000$ cases and $20{,}000$ controls, the equivalent quantitative sample size would be $26{,}667$. If we have an equal split of cases and controls, say $10{,}000$ of each, then the equivalent quantitative sample size would be exactly $20{,}000$.

4.2.3 Phenotypes

We collect summary statistic data for the largest available univariate GWAS on 19 traits. These traits are: Alzheimer's Disease (AD) (Lambert et al., 2013), Birth length (BL) (Valk et al., 2015), Bipolar Disorder (BPD) (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011), Birth weight (BW) (Horikoshi et al., 2013), Childhood obesity (CO) (The Early Growth Genetics (EGG) Consortium, 2012), College Yes/No (CY) (Rietveld et al., 2013), Extreme hip-waist ratio (EHW) (Berndt et al., 2013), Fasting glucose (GLU) (Dupuis et al., 2010), High-density lipoprotein (HDL) (Global Lipids Genetics Consortium, 2013), Height (Wood et al., 2014), Infant head circumference (IHC) (Taal et al., 2012), Low-density lipoprotein (LDL) (Global Lipids Genetics Consortium, 2013), Major Depressive Disorder (MDD) (Ripke et al., 2013), Obesity class 1: BMI ≥ 30 kg/m^2 (OC1), Obesity class 2: BMI ≥ 35 kg/m^2 (OC2), Obesity class 3: BMI ≥ 40 kg/m^2 (OC3) (Berndt et al., 2013), Schizophrenia (SCZ) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), Ever/never smoked (SM) (The Tobacco and Genetics Consortium, 2010) and Triglycerides (TG) (Global Lipids Genetics Consortium, 2013). We perform multi-trait summary statistic GWAS on all 19 traits, as well as subsets of the traits as described below.

4.2.4 Correlated-set analyses

In addition to performing multi-trait analyses on all traits jointly, we perform multi-trait analyses on sub-sets of correlated traits. From our comparison study in **Chapter 3**, we observed that the summary statistic GWAS methods could lose power when many traits are analysed jointly, due to increasing the degrees of freedom. All real data analyses, apart from those performed on all GWAS results jointly, were performed on groups of phenotypes with *a priori* known genetic correlations as established by the application of LD Score regression (B. Bulik-Sullivan et al., 2015). The groups of correlated traits to which the multi-trait methods are applied are listed in **Table 7**.

Correlated-sets	No. Traits
AD, CY	2
LDL, TG	2
BPD, MDD, SCZ	3
EHW, HDL, TG	3
HDL, LDL, TG	3
BL, BW, Height, IHC	4
BPD, CY, MDD, SCZ	4
CO, OC1, OC2, OC3	4
CY, EHW, HDL, LDL, TG	5
CY, OC1, OC2, OC3, SM	5
GLU, HDL, OC1, OC2, OC3	5
BW, GLU, HDL, OC1, OC2, OC3	6
CO, Height, HDL, OC1, OC2, OC3	6
AD, BPD, CY, HDL, OC1, OC2, OC3, TG	8
CO, CY, GLU, HDL, OC1, OC2, OC2, SM	8
CY, EHW, GLU, HDL, OC1, OC2, OC3, TG	8
All phenotypes	19

Table 7. Details of the correlated-sets of traits to which the MetaHom, MetaHet and Cattle methods are applied.

4.2.5 Summary statistics

Beta coefficients, standard errors and *P*-values for all SNPs genome-wide are obtained from the largest publicly available GWAS on each phenotype (stored, for example, here: https://www.med.unc.edu/pgc/results-and-downloads). These are used to compute the corresponding t-statistics to perform multi-trait GWAS using the MetaHom, MetaHet and Cattle methods (other methods not applied based on simulation results; see **4.3.1**) as described above. Chromosome and physical position of all SNPs are made consistent across data sets by assigning all to human genome build hg19 using the UCSC Genome Browser, and only those SNPs present across all 19 GWAS data sets are analysed. Across all 19 traits there are complete summary data for $838{,}294$ SNPs. For quantitative traits we use the sample size as reported in the individual studies, and for case/control phenotypes we convert to an effective sample size as given in **Equation 3**. For each group of traits to be analysed, we extract the corresponding t-statistics across all SNPs, and perform the analyses on these subsets.

4.2.6 Independent and novel associations

In order to compare the performance of the multi-trait analysis results to those of existing GWAS, the results are thinned according to linkage disequilibrium between SNPs to highlight only independent genome-wide significant associations. By chromosome, we used a 500kb window centred on each genome-wide significant SNP to remove correlated SNPs, retaining the SNP in the region with the smallest *P*-value. Significant associations were deemed to be novel in a multi-trait analysis if no SNP within the 500kb centred window on the multi-trait SNP had an association *P*-value below 5×10^{-8} in any of the relevant single-trait GWAS. We repeat this

procedure for all analyses conducted, establishing the subset of novel independent associations identified by each multi-trait method.

4.3 Results

4.3.1 Comparison of summary statistic GWAS methods

We begin by performing a comparison between the summary statistic based multi-trait GWAS methods MetaHom and MetaHet, which we have developed by extending previous methods to make them suitable for a mixture of continuous and case/control phenotype data and for negative genetic correlations, and the Cattle method (see **Materials and Methods**). In addition we include the min-*P* and TATES methods, which adjust univariate *P*-values, and CCA (mv-PLINK), which performs multi-trait analyses on individual-level genotype-phenotype data. We focus here on methods that perform SNP-by-SNP testing because, while there are methods for assessing pleiotropy at a locus-level (Pickrell et al., 2016), we wish to reveal single variant pleiotropic effects and produce results that are directly comparable to GWAS results on single phenotypes.

Figure 35 and **Figure 36** illustrate the results of comparing the multi-trait methods (TATES (van der Sluis et al., 2013), min-*P* (O'Reilly et al., 2012), MetaHom, MetaHet, Cattle (Bolormaa et al., 2014), CCA (mv-PLINK) (Ferreira and Purcell, 2009)) across a range of simulation scenarios. This comparison study, focusing on summary statistic methods, captures the most important aspects of method performance according to a more general multi-trait comparison study that we performed in **Chapter 3**. **Figure 35** shows the results of our simulations on two traits for scenario S1 (see **Chapter 2**), where the genetic effects on each trait are both the same

(**Figure 35a**), of different magnitudes (**Figure 35b**), and where only one trait is affected by the genetic variant (**Figure 35c**). We see that for homogeneous genetic effects (**Figure 35a**), as expected, MetaHom is the best performing method, while performing comparably poorly for the heterogeneous genetic effect simulations of **Figure 35c.** In contrast, MetaHet performs more consistently across the different scenarios, and the Cattle method performs comparably to the individual-level method CCA. Slight inflation is observed for the univariate methods min-*P* and TATES for large positive and negative correlations, while the other methods perform as expected under the null (**Figure 35d**). Further details of these simulations are provided in **Chapter 2** and **Chapter 3**.

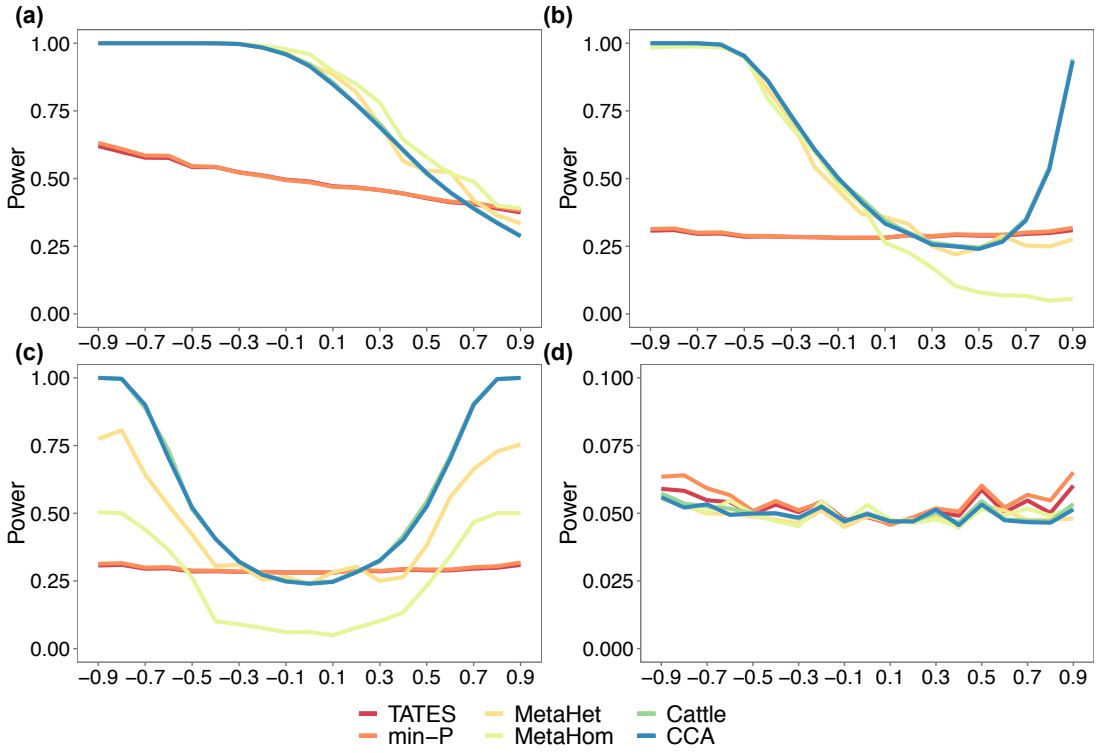

Figure 35. Simulation plots for scenario S1 with two phenotypes (as described in **Chapter 2**) for the summary statistic methods TATES, min-*P*, MetaHet, MetaHom and Cattle, as well as the individual-level method CCA (mv-PLINK): **(a)** same genetic effects, **(b)** different magnitude of genetic effects, **(c)** only one phenotype affected, **(d)** null effects.

The simulations of **Figure 36** are under the assumption that the genetic effects and phenotypic correlations are reflective of each other. Here the phenotypes are

simulated to be highly correlated if the genetic effects are of the same magnitude ($r = 0.6$), moderately correlated if the effects are of different magnitudes ($r = 0.2$), and have a low correlation if there is only one trait influenced by the genetic variant ($r = 0.05$). For two traits, three genetic effects vectors are simulated, and for four or more traits, 10 genetic effect vectors are simulated (see **Table 2** of **Chapter 2**). We observe that the univariate methods min-*P* and TATES consistently have the greatest performance. However as noted earlier, these methods are unable to achieve novel findings due to only performing adjustments on the univariate *P*-values. These methods both perform as expected since when the genetic effects are reflective of the phenotypic correlations, there is no gain in power from analysing the traits jointly (as in the other methods) as there is less potential for additional residual variance explained. In these scenarios, these results suggest that there would be no gain in performing a multi-trait analysis over the standard univariate approach, and thus if there is a causal effect it will have already been identified in the single-trait study. Out of the other methods, MetaHom performs well in the more pleiotropic scenarios, for example **Figure 36b** for genetic effect vector $\boldsymbol{v_1}$, where the magnitude of genetic effects are the same, and $\boldsymbol{v_5}$ where the genetic effect on three quarters of the traits is 0.5% and 0.1% on the remaining quarter of traits. However, MetaHom performs poorly for the least pleiotropic scenarios, for example **Figure 36d** for genetic effect vector $\boldsymbol{v_4}$, where the genetic variant is only affecting a quarter of the traits. The performance of MetaHet either matches the performance of the Cattle and CCA methods, or exceeds them across the different scenarios.

We choose to highlight these simulation results in particular as they expose both the similarities and differences in method performance across different scenarios, enabling us to gain insight into how these methods should perform on real data. The results highlight sensitivity of method performance to modelling scenario, but also

show that irrespective of scenario, TATES and min-*P* have matching performance, as do Cattle and CCA (mv-PLINK). MetaHom or MetaHet is typically the best-performing multi-trait method under pleiotropic and non-pleiotropic scenarios, respectively, and thus we expect the application of both methods to provide greatest discovery overall.

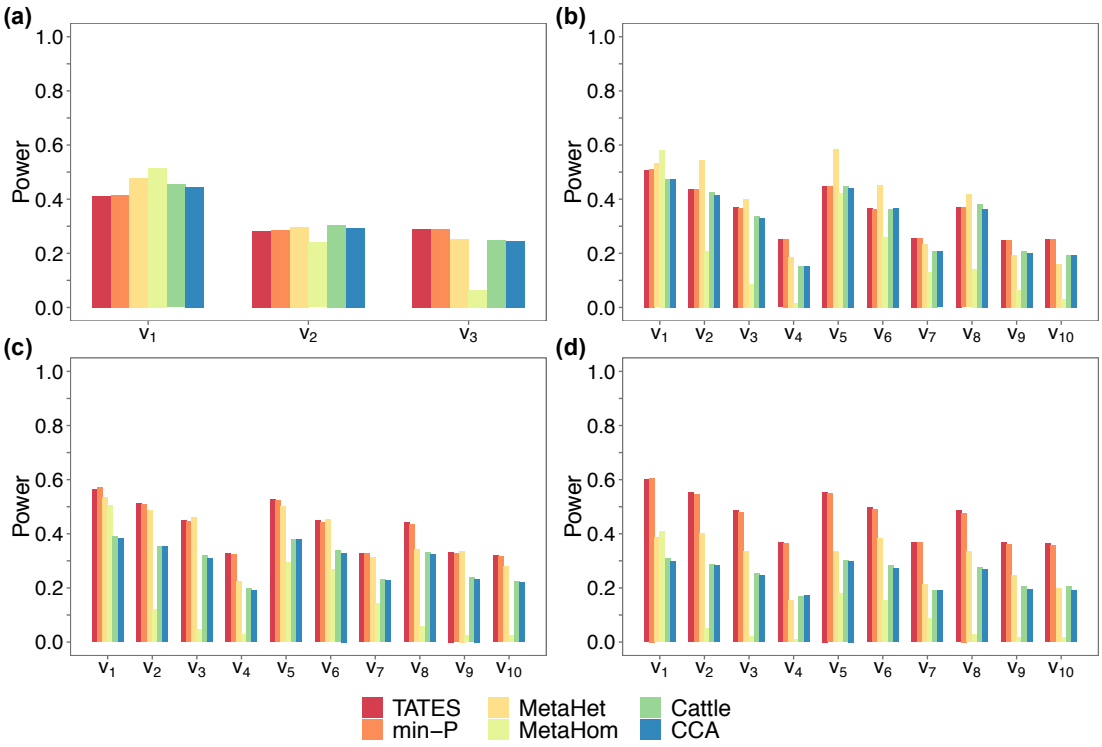

Figure 36. Simulation plots for scenario S3 (as described in **Chapter 2**) for the summary statistic methods TATES, min-*P*, MetaHet, MetaHom and Cattle, as well as the individual-level method CCA (mv-PLINK) for **(a)** 2, **(b)** 4, **(c)** 8 and **(d)** 12 phenotypes.

4.3.2 Summary statistic GWAS

Based on the results of the simulation study, we applied MetaHom and MetaHet to publicly available GWAS summary statistic data on 19 traits (see **Materials and Methods).** We also applied the Cattle method to the same data to represent the performance of the other methods and to verify the prediction from the simulations that MetaHom and MetaHet provide greatest overall discovery.

4.3.2.1 Correlated-set analyses

Our simulation study demonstrated that when these methods are applied to a large set of heterogeneous phenotypes the statistical power may be limited, and thus as well as application to all phenotypes we perform analyses on correlated-sets of traits. These correlated-sets were defined according to global estimates of genetic correlations from LD Score regression (see **Materials and Methods**) (B. K. Bulik-Sullivan et al., 2015). The genetic correlations from LD Score regression quantify the shared genetic aetiology between traits; these provide an indication of the pairs of phenotypes that have pleiotropic genetic effects. Analysing sets of homogenous traits with respect to genetic effects should maximise the power of the MetaHom method. **Table 8** displays the number of novel genome-wide significant SNPs identified for each of the 16 correlated-set analyses performed and for the analysis across all 19 traits jointly.

Phenotypes	MetaHom	MetaHet	Cattle
AD, BPD, CY, HDL, OC1, OC2, OC3, TG	30	37	18
AD, CY	1	1	0
All phenotypes	48	12	39
BL, BW, Height, IHC	68	12	10
BPD, CY, MDD, SCZ	10	20	6
BPD, MDD, SCZ	5	7	5
BW, GLU, HDL, OC1, OC2, OC3	10	12	7
CO, CY, GLU, HDL, OC1, OC2, OC2, SM	14	17	8
CO, Height, HDL, OC1, OC2, OC3	28	28	18
CO, OC1, OC2, OC3	2	0	0
CY, EHW, GLU, HDL, OC1, OC2, OC3, TG	33	39	18
CY, EHW, HDL, LDL, TG	31	44	44
CY, OC1, OC2, OC3, SM	1	1	2
EHW, HDL, TG	37	32	25
GLU, HDL, OC1, OC2, OC3	12	10	5
HDL, LDL, TG	38	45	50
LDL, TG	22	20	26
Total	**390**	**337**	**281**
Total unique	**244**	**169**	**151**

Table 8. Number of independent, novel genome-wide significant associations for the analyses of all phenotype subsets using the MetaHom, MetaHet and Cattle methods.

As expected from the simulations, MetaHom and MetaHet are the best-performing methods, and so hereafter we refer only to their results. We observe that MetaHom performs best for the analyses of traits where we may expect greatest pleiotropic genetic effects, for example, BL, BW, Height, and IHC. Interestingly, in the analysis of all 19 traits, MetaHom is again the best performing method, with MetaHet performing comparatively poorly. This may be due to the sub-setting procedure of MetaHet; in the analysis of all 19 traits, many subsets are tested for association with each SNP based on their univariate test statistics. While this can lead to increased power when only a subset of the analysed traits are associated with the SNP, it also greatly increases the degrees of freedom of the MetaHet test statistic, leading to a reduction in power overall. For the group of traits considered here, there are likely to be pleiotropic effects, for example between the lipids (HDL, LDL and TG) and obesity

measures (OC1, OC2 and OC3), as well as between the anthropometric traits such as height, birth length (BL), birth weight (BW) and infant head circumference (IHC), suggesting that MetaHom will perform well for the joint analysis of these traits.

In total, MetaHom identifies 390 independent, novel SNP-phenotype associations. Novelty is defined here according to there being no genome-wide significant associations in any of the univariate GWAS of the traits analysed, within a 500kb window centered on each SNP (see **Materials and Methods**). For example, one such SNP is rs1473886 at ~20.4Mb on chromosome 2, identified by the MetaHom correlated-set analysis of HDL, LDL and TG, which has a MetaHom association $P = 1.11 \times 10^{-19}$; thus, there are no genome-wide significant results between ~20.15Mb and ~20.65Mb in any of the independent GWAS on HDL, LDL and TG conducted by the Global Lipids Genetics Consortium (Global Lipids Genetics Consortium, 2013). A summary of the joint association *P*-value and the univariate *P*-values for this SNP is given in **Table 9**, and **Figure 37** illustrates the association results for HDL, LDL, TG and MetaHom in the region of this finding, highlighting the boost in statistical power achieved by combining the separate GWAS into a joint analysis.

Analysis	***P*-value**
HDL GWAS	1.14×10^{-5}
LDL GWAS	4.26×10^{-4}
TG GWAS	1.88×10^{-5}
Multi-trait GWAS of HDL, LDL and TG	1.11×10^{-19}

Table 9. Univariate *P*-values for the MetaHom top hit, rs1473886, from the univariate GWAS of HDL, LDL and TG, as well as the joint association *P*-value.

We see that the univariate *P*-values for rs1473886 are around the genome-wide suggestive significance threshold of 1×10^{-5}, and that the signal has been boosted to genome-wide significant in the multi-trait analysis.

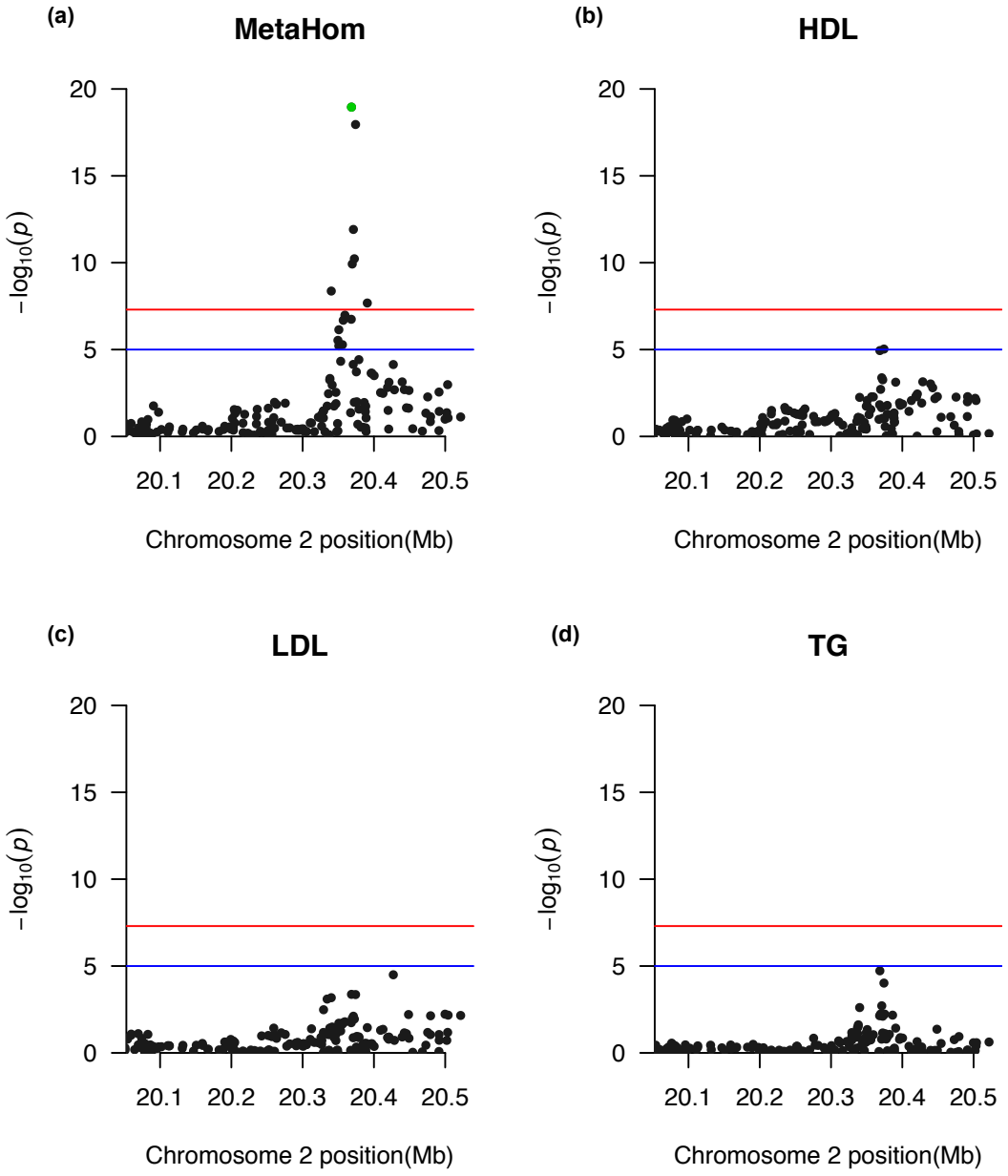

Figure 37. Manhattan plot for the MetaHom association results for chromosome 2 where the top hit from the MetaHom analysis of HDL, LDL and TG (rs1473886; $P = 1.11 \times 10^{-19}$) is located. Manhattan plots for the univariate GWAS on HDL, LDL and TG for the same region.

4.3.2.2 Multiple-testing correction

To adjust for multiple testing, for each method we calculated the correlation between the test statistics across all 16 correlated-set analyses and the analysis of all 19 traits. We then applied a Nyholt correction (Nyholt, 2004) to calculate the number of independent tests performed. **Table 10** details the number of independent tests

performed for each method and the adjusted significance threshold, which was also further adjusted to account for the application of three methods.

Method	Number of Tests	Adjusted Threshold
MetaHom	13.16	1.27×10^{-9}
MetaHet	12.57	1.33×10^{-9}
Cattle	12.60	1.32×10^{-9}

Table 10. Number of independent tests performed across the 17 analyses for each method, and the corresponding adjusted significance thresholds.

Table 11 provides the adjusted number of novel, independent associations after correcting for multiple testing.

Phenotypes	MetaHom	MetaHet	Cattle
AD, BPD, CY, HDL, OC1, OC2, OC3, TG	7	12	4
AD, CY	0	0	0
All phenotypes	11	2	8
BL, BW, Height, IHC	1	0	2
BPD, CY, MDD, SCZ	0	0	0
BPD, MDD, SCZ	0	2	0
BW, GLU, HDL, OC1, OC2, OC3	2	4	1
CO, CY, GLU, HDL, OC1, OC2, OC2, SM	2	6	1
CO, Height, HDL, OC1, OC2, OC3	8	9	4
CO, OC1, OC2, OC3	1	0	0
CY, EHW, GLU, HDL, OC1, OC2, OC3, TG	11	14	6
CY, EHW, HDL, LDL, TG	11	17	11
CY, OC1, OC2, OC3, SM	0	0	0
EHW, HDL, TG	16	8	4
GLU, HDL, OC1, OC2, OC3	2	3	0
HDL, LDL, TG	13	18	11
LDL, TG	4	3	4
Total	**89**	**98**	**56**
Total unique	**50**	**45**	**34**

Table 11. Number of independent, novel hits at the multiple-testing significance threshold for the analyses of all phenotype subsets using the MetaHom, MetaHet and Cattle methods.

After adjusting for multiple testing, MetaHom still produces the largest number of novel findings overall. The Cattle method performs worse than both the MetaHom and MetaHet methods. **Figure 38** provides a visual representation of the number of novel, independent genome-wide significant associations for the MetaHom and MetaHet methods across the correlated-set analyses, after adjusting for multiple testing. **Table 12** and **Table 13** provide details of the genome-wide significant associations (after multiple testing correction) for the multi-trait analyses of HDL, LDL and TG for the MetaHom and MetaHet methods respectively.

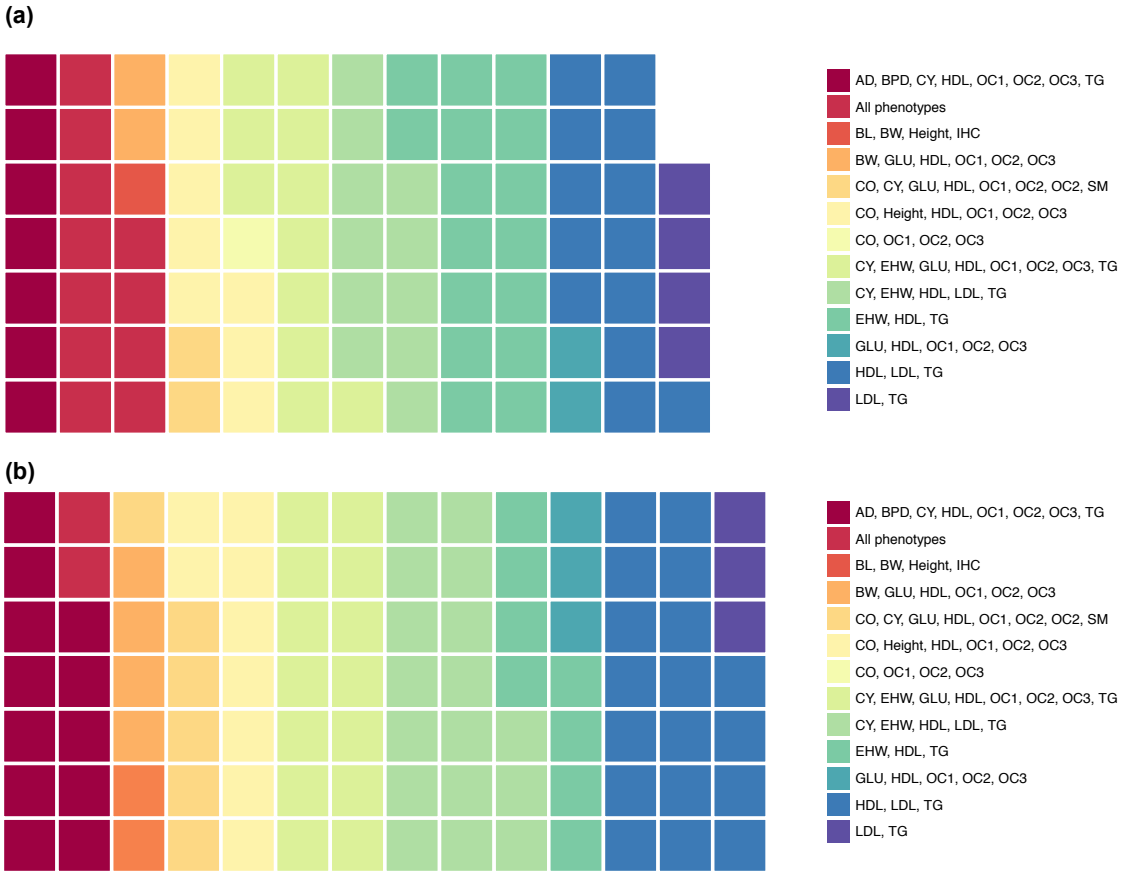

Figure 38. Waffle plots displaying the proportions of the multiple-testing adjusted novel, independent associations (1 square = 1 SNP) for each correlated-set analysis for the **(a)** MetaHom and **(b)** MetaHet methods.

SNP	CHR	Position (Mb)	*P*-value
rs1473886	2	20.4	1.11×10^{-19}
rs10904908	10	17.3	9.42×10^{-15}
rs3779500	7	10.7	9.51×10^{-13}
rs7033354	9	16.9	2.78×10^{-12}
rs6505081	17	26.7	3.74×10^{-12}
rs1997243	7	1.08	6.25×10^{-12}
rs7947951	11	13.4	1.13×10^{-11}
rs1198433	1	23.8	1.16×10^{-10}
rs4933744	10	94.6	2.05×10^{-10}
rs12539895	7	107.1	2.30×10^{-10}
rs12137896	1	26.7	3.72×10^{-10}
rs1382567	8	11.4	5.86×10^{-10}
rs2936507	8	6.61	1.13×10^{-9}

Table 12. Novel, independent genome-wide significant SNPs (after multiple testing correction) for the multi-trait analysis of HDL, LDL and TG using the MetaHom method.

SNP	CHR	Position (Mb)	*P*-value
rs1473886	2	20.4	9.25×10^{-19}
rs10904908	10	17.3	6.02×10^{-14}
rs3779500	7	106.8	5.39×10^{-12}
rs7033354	9	16.9	1.53×10^{-11}
rs6505081	17	26.7	2.04×10^{-11}
rs1997243	7	1.08	2.98×10^{-11}
rs7947951	11	13.4	5.97×10^{-11}
rs2862954	10	101.9	9.53×10^{-11}
rs2287623	2	169.8	2.13×10^{-10}
rs10928512	2	135.5	2.57×10^{-10}
rs10513801	3	185.8	4.23×10^{-10}
rs1198433	1	23.8	5.73×10^{-10}
rs4895441	6	135.4	7.94×10^{-10}
rs10861661	12	107.2	9.61×10^{-10}
rs4933744	10	94.6	9.97×10^{-10}
rs6968554	7	17.3	1.09×10^{-9}
rs12539895	7	107.1	1.11×10^{-9}
rs6901147	6	52.5	1.23×10^{-9}

Table 13. Novel, independent genome-wide significant SNPs (after multiple testing correction) for the multi-trait analysis of HDL, LDL and TG using the MetaHet method.

The top novel hit from the analysis of HDL, LDL and TG across both the MetaHom and MetaHet methods, rs1473886 on chromosome 2, was identified in a recent univariate analysis of total cholesterol from the Global Lipids Genetics Consortium ($P = 5.06 \times 10^{-10}$) (Global Lipids Genetics Consortium, 2013). Of the MetaHet hits that were not identified by the MetaHom method a range of other traits have been found to show the greatest level of association in previous studies. The association of rs2862954 with fatty liver disease and alanine aminotransferase (ALT) levels ($P = 3.03 \times 10^{-10}$) (Feitosa et al., 2013) is the most significant currently recorded for that SNP in the *PhenoScanner* tool (Staley et al., 2016), which provides an online catalog of all published GWAS results. Among other associated traits with the MetaHet hits are BMI for rs10513801 ($P = 1.94 \times 10^{-23}$) (Locke et al., 2015), as well as coffee consumption ($P = 7.00 \times 10^{-15}$) (Cornelis et al., 2015) and blood metabolite levels ($P = 9.00 \times 10^{-14}$) (Shin et al., 2014) for rs6968554. Several known total cholesterol loci were identified in the MetaHom analysis of LDL, HDL and TG, such as *ASAP3* (chr 1: 23.8Mb), *GPR146* (chr7: 1.08Mb) and *VIM-CUBN* (chr10: 17.3Mb) (Global Lipids Genetics Consortium, 2013). The MetaHet analysis, in addition, identified the *ABCB11* (chr 2: 169.8Mb) and *HBS1L* (chr6: 135.4Mb) total cholesterol associated loci (Global Lipids Genetics Consortium, 2013).

These observations provide validation for our findings and approach by replicating known total cholesterol hits, as well as highlighting the genetic link between the analysed traits and other traits known to be associated with the identified SNPs, such as the lipids and fatty liver disease. Our results also provide evidence that analysing components of traits in a multivariate model could lead to additional power gains, as was achieved here by analysing the lipids in a joint model instead of only total cholesterol. Furthermore, these findings indicate that multi-trait methods could not only be applied to gain additional power for identifying SNP-phenotype associations,

but could also provide useful information for identifying sources of shared genetic aetiology between traits.

In addition to the aim of identifying novel association signals by repurposing existing GWAS summary data, we also sought to compare how MetaHom and MetaHet perform in practice on real data. From these results, it would suggest that manually sub-setting the traits should be performed prior to combining them in a joint analysis, as done here using the LD score genetic correlations (B. Bulik-Sullivan et al., 2015), in order to yield greatest discovery.

4.3.3 Validation using CADD scores

To validate our findings, we used the CADD (Combined Annotation Dependent Depletion) score (Kircher et al., 2014). The CADD score quantifies how deleterious a SNP is likely to be based on estimated intensities of purifying selection across multiple functional annotation categories, and has been shown to be correlated with pathogenicity and disease severity (Kircher et al., 2014). We compared the CADD scores of both novel and all MetaHom and MetaHet hits with those of all independent genome-wide significant hits from the latest GWAS on schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), height (Wood et al., 2014), HDL and LDL (Global Lipids Genetics Consortium, 2013), and compared each of these with a set of randomly selected SNPs (see **Table 14**).

SNPs	CADD Score	No. SNPs	*P*-value vs. Random
MetaHom novel	4.39	236	7.45×10^{-4}
MetaHom all	4.55	1807	2.19×10^{-22}
MetaHet novel	4.58	165	1.23×10^{-3}
MetaHet all	4.50	1542	1.51×10^{-18}
SCZ GWAS	4.35	154	1.04×10^{-2}
HDL GWAS	4.97	123	4.86×10^{-3}
LDL GWAS	5.21	105	4.46×10^{-3}
Height GWAS	4.56	721	7.06×10^{-11}
Random SNPs	3.38	5000	-

Table 14. Mean CADD scores for the MetaHom and MetaHet total independent hits and novel, independent hits, as well as for independent GWAS hits for SCZ, HDL, LDL and Height and a random set of SNPs.

The CADD scores of the MetaHom and MetaHet results were similar to those of the published GWAS and significantly higher than those of the random SNPs (novel MetaHom vs. random: $P = 7.45 \times 10^{-4}$; novel MetaHet vs. random $P = 1.23 \times 10^{-3}$; all MetaHom vs. random $P = 2.19 \times 10^{-22}$; all MetaHet vs. random $P = 1.51 \times 10^{-18}$). These results, combined with those of the simulations showing that MetaHom and MetaHet have appropriate type 1 error (**Figure 35d**), strongly support the validity of our findings.

Table 15 provides the mean CADD scores for the novel, independent genome-wide significant SNPs and the independent genome-wide significant SNPs, after adjusting for multiple testing.

SNPs	CADD score	No. SNPs	*P*-value vs. Random
MetaHom novel	5.04	48	0.0322
MetaHom all	4.70	1271	3.97×10^{-20}
MetaHet novel	5.30	43	0.0342
MetaHet all	4.60	1027	7.72×10^{-15}

Table 15. Mean CADD scores for the MetaHom and MetaHet total independent and novel, independent hits after adjusting for multiple testing.

We observe that the mean CADD scores for the MetaHom and MetaHet SNPs after adjusting for multiple testing are significantly greater than the random SNPs, for both the independent and novel, independent SNPs.

4.3.4 Boost to univariate signals

As well as identifying novel genome-wide significant findings through performing multi-trait analyses on GWAS summary statistics, we observe that the univariate signals of previously identified genetic associations are typically boosted. **Table 16** details the SNPs that were found to be genome-wide significant in the univariate LDL GWAS, and were also identified by the MetaHom method in the HDL, LDL and TG correlated-set analysis. Here we observe that the univariate *P*-values are generally boosted by the multi-trait analysis, with quite dramatic boosts observed for some SNPs. For example, rs4587594 on chromosome 1 has a univariate association *P*-value of 1.63×10^{-32} from the LDL GWAS, and achieves a multi-trait *P*-value of 1.21×10^{-125} from the MetaHom analysis. This same SNP has a univariate *P*-value of 3.50×10^{-82} from the TG GWAS, and was not identified as genome-wide significant in the HDL GWAS.

SNP	CHR	Position (Mb)	univariate *P*-value	multi-trait *P*-value
rs11206510	1	55.5	2.38×10^{-53}	2.12×10^{-34}
rs207145	1	55.8	6.19×10^{-18}	2.36×10^{-8}
rs2807834	1	221.0	1.19×10^{-15}	2.07×10^{-31}
rs4587594	1	63.1	1.63×10^{-32}	1.21×10^{-125}
rs518076	1	110.1	5.60×10^{-14}	3.37×10^{-12}
rs558971	1	234.9	5.56×10^{-24}	1.25×10^{-12}
rs660240	1	109.8	9.00×10^{-265}	1.03×10^{-162}
rs2250802	10	113.9	3.94×10^{-13}	5.98×10^{-40}
rs10832962	11	18.7	6.62×10^{-14}	1.97×10^{-9}
rs1535	11	61.6	7.77×10^{-41}	5.89×10^{-126}
rs4937122	11	126.2	1.81×10^{-20}	2.70×10^{-19}

rs964184	11	116.6	2.01×10^{-26}	1.02×10^{-154}
rs10774625	12	111.9	8.31×10^{-11}	4.72×10^{-21}
rs11066320	12	112.9	1.57×10^{-9}	1.58×10^{-16}
rs17630235	12	112.6	5.61×10^{-11}	1.17×10^{-18}
rs735396	12	121.4	5.36×10^{-16}	8.66×10^{-14}
rs4942486	13	33.0	2.26×10^{-11}	7.57×10^{-11}
rs8017377	14	24.9	2.52×10^{-15}	3.99×10^{-8}
rs1532624	16	57.0	1.84×10^{-26}	$< 5 \times 10^{-324}$
rs217181	16	72.1	1.13×10^{-25}	8.79×10^{-20}
rs3809868	17	45.8	4.30×10^{-10}	5.06×10^{-23}
rs6504872	17	45.4	3.48×10^{-13}	7.89×10^{-23}
rs10460181	19	45.1	2.25×10^{-28}	4.99×10^{-35}
rs11881315	19	10.9	7.75×10^{-50}	3.28×10^{-29}
rs16996148	19	19.7	1.97×10^{-45}	7.99×10^{-54}
rs2075650	19	45.4	1.72×10^{-214}	1.77×10^{-174}
rs2228603	19	19.3	4.43×10^{-44}	6.06×10^{-48}
rs6511720	19	11.2	3.85×10^{-262}	6.60×10^{-132}
rs676388	19	49.2	1.31×10^{-11}	2.12×10^{-10}
rs10195252	2	165.5	3.81×10^{-8}	3.10×10^{-23}
rs4988235	2	136.6	3.22×10^{-11}	8.81×10^{-12}
rs6544713	2	44.1	4.84×10^{-83}	1.85×10^{-39}
rs6730157	2	135.9	6.32×10^{-11}	4.90×10^{-10}
rs6739502	2	21.5	5.27×10^{-25}	9.44×10^{-17}
rs693	2	21.2	1.20×10^{-131}	7.36×10^{-129}
rs6016381	20	39.2	6.85×10^{-20}	2.66×10^{-16}
rs6102322	20	39.9	9.63×10^{-20}	1.13×10^{-13}
rs11709504	3	12.7	4.60×10^{-8}	1.36×10^{-12}
rs9875338	3	12.3	2.21×10^{-11}	2.45×10^{-14}
rs1501908	5	156.4	1.12×10^{-28}	3.49×10^{-27}
rs34358	5	75.0	8.77×10^{-31}	8.55×10^{-28}
rs3935470	5	74.4	1.59×10^{-27}	5.99×10^{-21}
rs6878990	5	74.7	5.83×10^{-63}	1.60×10^{-45}
rs11753995	6	160.6	4.06×10^{-21}	7.03×10^{-15}
rs1367211	6	161.1	7.23×10^{-9}	5.64×10^{-12}
rs2294261	6	16.1	6.57×10^{-17}	5.12×10^{-11}
rs868943	6	116.3	8.44×10^{-11}	9.61×10^{-14}
rs12670798	7	21.6	4.81×10^{-14}	4.25×10^{-8}
rs17725246	7	44.6	1.49×10^{-20}	1.13×10^{-12}
rs4722551	7	26.0	3.95×10^{-14}	5.72×10^{-28}
rs2126259	8	9.19	1.35×10^{-22}	1.59×10^{-55}
rs2326077	8	59.4	5.00×10^{-17}	6.00×10^{-17}

rs6982502	8	126.5	4.70×10^{-46}	2.58×10^{-120}
rs6993938	8	145.0	1.77×10^{-10}	4.09×10^{-10}
rs1883025	9	107.7	6.14×10^{-11}	4.48×10^{-104}
rs579459	9	136.2	2.42×10^{-44}	9.24×10^{-52}

Table 16. *P*-values for the genome-wide significant SNPs in the univariate GWAS on LDL, and the joint association *P*-values for the joint analysis of HDL, LDL and TG using the MetaHom method.

We do not observe a boost in all univariate *P*-values, however, suggesting that these SNPs may not have a pleiotropic effect on the analysed traits.

Furthermore, we observe that SNPs that reached suggestive level of significance in the univariate GWAS often become genome-wide significant in the multi-trait analysis. **Table 17** details the SNPs from the univariate LDL GWAS with $1 \times 10^{-5} \leq P < 5 \times 10^{-8}$ that were found to be genome-wide significant in the multi-trait GWAS of HDL, LDL and TG using the MetaHom method; similarly for the MetaHet method in **Table 18**.

SNP	CHR	Position (Mb)	univariate *P*-value	multi-trait *P*-value
rs17162330	1	27.2	8.28×10^{-6}	2.74×10^{-18}
rs2095403	1	62.9	9.27×10^{-6}	9.57×10^{-10}
rs1260326	2	27.7	1.51×10^{-7}	2.44×10^{-267}
rs2287623	2	169.8	5.40×10^{-8}	2.13×10^{-10}
rs13315871	3	58.4	4.72×10^{-7}	3.42×10^{-9}
rs6831256	4	3.5	9.07×10^{-7}	5.77×10^{-14}
rs3891175	6	32.6	7.53×10^{-6}	8.62×10^{-13}
rs2814993	6	34.6	2.87×10^{-7}	5.68×10^{-18}
rs11755266	6	35.2	1.98×10^{-6}	1.14×10^{-13}
rs6901147	6	52.5	7.72×10^{-7}	1.23×10^{-9}
rs4895441	6	135.4	7.28×10^{-6}	7.94×10^{-10}
rs4921914	8	18.3	1.92×10^{-7}	1.52×10^{-18}
rs7033354	9	16.9	1.42×10^{-6}	1.53×10^{-11}
rs10904908	10	17.3	3.14×10^{-7}	6.02×10^{-14}
rs4933744	10	94.6	2.55×10^{-6}	9.97×10^{-10}
rs1351452	11	116.9	4.16×10^{-7}	2.15×10^{-89}
rs7117842	11	122.5	7.56×10^{-7}	1.13×10^{-20}
rs4765219	12	124.4	8.90×10^{-6}	2.46×10^{-19}
rs7193549	16	71.6	1.54×10^{-6}	1.34×10^{-8}
rs2277862	20	34.2	1.30×10^{-6}	4.36×10^{-8}

Table 17. Suggestive hits from the univariate LDL GWAS that were found to be genome-wide significant in the multi-trait analysis of HDL, LDL and TG using the MetaHom method.

SNP	CHR	Position (Mb)	univariate *P*-value	multi-trait *P*-value
rs17162330	1	27.2	8.28×10^{-6}	3.36×10^{-19}
rs1260326	2	27.7	1.51×10^{-7}	1.65×10^{-62}
rs11688816	2	63.1	2.30×10^{-7}	8.32×10^{-9}
rs6430552	2	135.6	1.15×10^{-6}	3.43×10^{-8}
rs2287623	2	169.8	5.40×10^{-8}	1.57×10^{-8}
rs13315871	3	58.4	4.72×10^{-7}	4.51×10^{-9}
rs6831256	4	3.47	9.07×10^{-7}	6.86×10^{-13}
rs3891175	6	32.6	7.53×10^{-6}	1.45×10^{-13}
rs3800457	6	34.7	2.30×10^{-7}	1.06×10^{-14}
rs11755266	6	35.2	1.98×10^{-6}	2.97×10^{-9}
rs6901147	6	52.5	7.72×10^{-7}	8.09×10^{-9}
rs4921914	8	18.3	1.92×10^{-7}	1.83×10^{-19}
rs7033354	9	16.9	1.42×10^{-6}	2.78×10^{-12}
rs10904908	10	17.3	3.14×10^{-7}	9.42×10^{-15}
rs4933744	10	94.6	2.55×10^{-6}	2.05×10^{-10}
rs7117842	11	122.5	7.56×10^{-7}	1.33×10^{-12}
rs10744777	12	112.2	1.56×10^{-7}	3.98×10^{-11}
rs4765219	12	124.4	8.90×10^{-6}	2.86×10^{-20}
rs7193549	16	71.6	1.54×10^{-6}	3.73×10^{-9}

Table 18. Suggestive hits from the univariate LDL GWAS that were found to be genome-wide significant in the multi-trait analysis of HDL, LDL and TG using the MetaHet method.

4.4 Discussion

By re-analysing existing GWAS results, combining information across traits, we have identified a large number of novel genotype-phenotype associations. While power can also be increased by additional sample size, this study provides a proof-of-principle that no matter what the present sample sizes of GWAS, novel variants can be discovered by performing multi-trait analyses on GWAS summary data. This can be particularly beneficial for traits that are highly heterogeneous or for which sample collection is challenging, since their independent GWAS may have produced few findings and thus targets for further biological investigation or drug development. The variants with the strongest signals from our cross-trait analyses may point to causal variants that act as 'hubs' of phenotypic importance, deserving special biological investigation, which may unravel the mechanisms underlying the shared aetiology between the traits.

Our results have demonstrated that performing multi-trait analyses of GWAS summary statistics can boost univariate signals. This may provide the increase in power required to identify genetic variants associated with traits that have few known genetic determinants to date. The resource of publicly available summary data is continually growing and there is now an online tool, *PhenoScanner* (Staley et al., 2016), that collates summary statistics across a wide range of traits, further facilitating these types of analyses. The traits considered in this study could be considered to be rather homogenous in respect to genetic effects, thus further multi-trait analyses should be performed to gain further insight into the interaction between genetic variation and correlated sets of traits. Here we also applied both the MetaHom and MetaHet methods, as our simulations suggested that a combination of the two methods produces the most novel findings when the causal genetic

relationships are not known. However, we did not test this inference formally and applying two methods increases the multiple testing burden, providing the motivation for methodology development into the identification of the type of causal genetic relationships that typically exist, so that the most appropriate method can be adopted.

A detailed evaluation of our results can provide more general insights into the genetic aetiology underlying correlated traits; strong results for MetaHom compared to MetaHet, which we observe for most of the correlated-sets of traits here, are indicative of relatively uniform effects across the tested traits. The set of MetaHom and MetaHet GWAS results produced here, in particular the former, provide the field with a 'pleiotropic GWAS' resource, both relating to a wide range of phenotypes and more focused subsets. This can act as a unified GWAS results set, capturing effects across multiple traits into single measures of association. Such a resource is useful for testing the enrichment of, for example, regulatory features, genomic aberrations or putative selected loci, against a single set of GWAS signals. We expect multi-trait GWAS scans to be continually updated as univariate GWAS of larger sizes are performed and publicly released, and we believe that they will aid in our understanding of specific phenotypes and in the biological processes contributing to the ubiquitous correlation between human phenotypes.

5. Building a multi-trait predictive model of Major Depressive Disorder

Major Depressive Disorder (MDD) is a common, polygenic disorder that is estimated to affect more than 350 million people worldwide, and is the leading cause of disability in the world ('WHO | Depression', Fact Sheet; Whiteford et al., 2013). MDD is characterised by low mood, loss of interest, feelings of guilt, reduced appetite, and disrupted sleep. While MDD has been shown to be highly heritable ($h^2 = 37\%$) (Sullivan et al., 2000), research efforts have discovered only a modest number of genetic markers associated with MDD (Ripke et al., 2013; Converge Consortium, 2015; Hyde et al., 2016) compared to other psychiatric disorders such as schizophrenia (SCZ), for which there are 108 independent known loci (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Genome-wide association studies (GWAS) have often been limited by small sample sizes and/or high heterogeneity in phenotype, resulting in low statistical power to detect causal associations. A recent study by 23andMe (Hyde et al., 2016) identified 4 independent loci from a discovery GWAS on $326{,}113$ individuals ($84{,}847$ cases). The 23andMe data were meta-analysed with the PGC MDD GWAS (Ripke et al., 2013), with this previous smaller GWAS failing to find any loci (Ripke et al., 2013). This demonstrates that due to the heterogeneous nature of MDD, extremely large sample sizes are required to achieve a well-powered study.

Neuroticism is a personality trait, with high levels of neuroticism characterised by anxiety, worry, loneliness, feelings of guilt and sensitivity, and heritability is estimated to be between 15% and 37% (Smith et al., 2016). Neuroticism is considered a continuous measure, with the severity determined by 12 component measures; a neuroticism score between 0 and 12 is computed from answers to the Eysenck

Personality Questionnaire - Revised Short Form (EPQ-R-S) (Eysenck et al., 1985). High neuroticism scores are correlated with MDD (Jylhä and Isometsä, 2006), and a genetic correlation between neuroticism and MDD has been established (Smith et al., 2016).

As MDD is a polygenic disorder, polygenic risk scores (PRS) are often constructed based on MDD GWAS results to quantify an individual's genetic risk for MDD. The resulting PRS is a weighted sum of the risk alleles for MDD carried by an individual across the genome (Purcell et al., 2009; Dudbridge, 2013; Euesden et al., 2015). This technique allows exploration into the genetic aetiology of disease, and is often used as a tool for predicting case/control status. While the PRS approach has been successful in other complex disorders (Purcell et al., 2009; Selzam et al., 2016), the low power of MDD GWAS to date has meant that PRS for MDD generally explain a relatively small amount of the variance in MDD diagnosis. Heterogeneity of the MDD phenotype could also provide an explanation for the poor predictive power of the MDD PRS.

In this chapter we build predictive models of MDD case/control status utilising the rich phenotyping available in the UK Biobank resource (http://www.ukbiobank.ac.uk/). Usually only PRS for MDD and SCZ are used as predictors of MDD (Ripke et al., 2013; Euesden et al., 2015). Here we build PRS on additional phenotypes that associate with MDD, such as component measures of neuroticism, obesity and whether college/university was attended. We build multivariate predictive models of MDD using these PRS predictors, choosing the most predictive model by variable selection techniques. Furthermore, we consider two-way interactions between these PRS variables to gain further insight into the genetic aetiology of MDD and its associated outcomes. Previous genetic interactions have been explored on a SNP-by-SNP basis; here we aim to capture broader genetic interaction using PRS as a

proxy for the overall genetic predisposition to the corresponding phenotype, which may present phenotypically even if not clinically.

5.1 Introduction

We aim to build predictive models for Major Depressive Disorder (MDD) that exploit multiple predictors, with an assumption that MDD is best predicted by a number of component measures that together capture different forms of this highly heterogeneous disorder. Given the complexity of MDD we also include interactions between the component measures as potential predictors. While we use polygenic risk scores (PRS) based on GWAS on MDD and schizophrenia from the Psychiatric Genomics Consortium (PGC) (Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), we build all other PRS, and train and test our predictive models, using UK Biobank data. GWAS are performed on the set of component measures in 80% of the UK Biobank data, a further 10% of the data is used to train the predictive models, and the final 10% of the data is used to validate the most predictive models. As well as the 12 components of neuroticism for which we build PRS, we also build PRS on obesity and whether college was attended. These PRS, together with the PRS for MDD and SCZ, are included as predictors in the models, and age, sex and 15 genetic principal components (to control for population structure) are included as control variables.

The first part of the chapter focuses on prediction of MDD from PRS alone. We first consider the main effect terms of the PRS, performing stepwise variable selection using both Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) on the model as follows:

$$MDD_{status} \sim MDD_{PRS} + SCZ_{PRS} + NEU1_{PRS} + NEU2_{PRS} + \ldots + NEU12_{PRS} + Obese_{PRS} + College_{PRS} + Age + Sex + PC_1 + PC_2 + \cdots + PC_{15}$$

where $NEU1_{PRS}$ is the PRS for the first component measure of neuroticism (here 'mood swings'), $NEU2_{PRS}$ is the PRS for the second component measure of neuroticism (here 'miserableness'), and so on.

Next we include interaction terms, exhaustively including all two-way interactions between each component PRS and age and sex, so that variable selection is applied to the model containing 153 interaction terms in addition to the 16 main effect terms and the control variables.

We also consider phenotype-only prediction models, which do not include genetic data in the form of PRS as predictors. We perform variable selection in the same way as for the PRS models (using both AIC and BIC) on both main effect and two-way interaction models, as detailed below.

The main effect model with phenotype predictors on which variable selection is performed is given by:

$$MDD_{status} \sim NEU_1 + NEU_2 + \ldots + NEU_{12} + Obese + College + Age + Sex$$

The two-way interaction model of the phenotype predictors on which variable selection is performed contains 120 interaction terms in addition to the 14 main effect terms and the age and sex control variables.

Finally, we build predictive models of MDD using both genetic (PRS) and phenotypic predictors. We build models using only the main effects of the PRS and phenotype predictors, as well as allowing interactions between the genetic predictors,

interactions between the phenotype predictors, and interactions between the genetic and phenotype predictors.

Below, we begin by detailing the materials and methods used in the study before describing the results.

5.2 Materials and Methods

The UK Biobank (Sudlow et al., 2015) is a population-based resource of approximately $502{,}000$ individuals aged between 39 and 73, for which recruitment took place between 2006 and 2010. The current release of genetic data (June 2015) includes genotype data on $152{,}249$ individuals. Data on a huge variety of health and trait measures have been collected as part of the UK Biobank, such as physical health outcomes, personality traits, and cognitive functioning, as well as oncology.

5.2.1 Phenotypes

Here we focus on MDD and its associated traits. In particular, for building polygenic risk scores (PRS) for prediction we focus on neuroticism due to its established genetic and phenotypic correlation with MDD (Smith et al., 2016; Kendler and Myers, 2010). In addition to using the neuroticism score as defined by the 12-point Eysenck Personality Questionnaire - Revised Short Form (EPQ-R-S) (Eysenck et al., 1985), we also consider these 12 questions as endo-phenotypes and perform analyses on these binary traits individually. These 12 endo-phenotypes are: mood swings (MS), miserableness (M), irritability (I), sensitivity (S), fed-up feelings (FU), nervous feelings (N), worrier (W), tense feelings (T), worry too long after embarrassment (WTL), suffer from nerves (SFN), loneliness (L) and guilty feelings (G). Details of the questions asked to UK Biobank participants are given in **Table 19**. In addition, we produce PRS on obesity, defined as having BMI ≥ 30 kg/m^2 (Yang et al., 2012), and on educational attainment (whether college/university was attended), due to the known relationships with MDD (Hung et al., 2015; Rivera et al., 2012; Bjelland et al., 2008).

Component	Abbr.	Question
Mood swings	MS	Does your mood often go up and down?
Miserableness	M	Do you ever feel 'just miserable' for no reason?
Irritability	I	Are you an irritable person?
Sensitivity	S	Are your feelings easily hurt?
Fed-up feelings	FU	Do you often feel 'fed-up'?
Nervous feelings	N	Would you call yourself a nervous person?
Worrier	W	Are you a worrier?
Tense feelings	T	Would you call yourself tense or 'highly strung'?
Worry too long	WTL	Do you worry too long after an embarrassing experience?
Suffer from nerves	SFN	Do you suffer from 'nerves'?
Loneliness	L	Do you often feel lonely?
Guilty feelings	G	Are you often troubled by feelings of guilt?

Table 19. Summary of the 12 neuroticism endo-phenotypes, the symbol we use to refer to them, and the question that was asked to UK Biobank participants to determine a diagnosis. Participants were assessed via a touchscreen questionnaire, and answered: yes, no, do not know, or prefer not to answer.

As some of the phenotypes under study do not currently have publicly available GWAS summary statistics, and due to potential problems with sample overlap in the UK Biobank, we split our data into training and test datasets. We use 80% of the UK Biobank as the training data in which to perform the discovery GWAS, required to compute polygenic risk scores, on the 15 traits detailed above. We use 10% of the UK Biobank as the test dataset for building the PRS and for building the predictive model of MDD. We retain a further 10% of the UK Biobank as a validation dataset in order to cross-validate the model(s) found to be most predictive of MDD in the test dataset.

5.2.2 Genotyping

Genotyping of the UK Biobank sample was performed by Affymetrix using the UK BiLEVE and UKB Axiom arrays, which are two customised microarrays that assay over $800{,}000$ variants. For further information on the genotyping process, see

http://www.ukbiobank.ac.uk/wpcontent/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf.

5.2.3 Genotype quality control

Preliminary quality control (QC) was performed by Affymetrix and the UK Biobank (see http://www.ukbiobank.ac.uk/wpcontent/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf). In addition, we performed study specific QC using PLINK-1.9 (Chang et al., 2015) as follows. We removed all markers with $MAF < 0.01$, and a genotype missingness threshold of 0.05 was applied. Hardy-Weinberg Equilibrium (HWE) was tested at a threshold of 5×10^{-8}. Individuals with genotype missingness greater than 0.1 were removed and sex discordant individuals were excluded from the analysis. We restrict our sample to unrelated individuals using a Kinship coefficient threshold of 0.05, and to those individuals with a homogeneous UK-dominated ancestry, as defined by a cross-section of 4-means clustering on the first two principal components from a principal components analysis on the full genome-wide data. We performed GWAS on non-imputed data, consisting of $498{,}104$ SNPs after QC.

5.2.4 MDD phenotyping

We defined MDD cases as any reported primary diagnosis of MDD from inpatient hospital records, identified by ICD10 subchapters F32 and F33. In addition, an individual was defined as a case for MDD if they reported a visit to a GP or psychiatrist for stress, anxiety or depression, and at least one period of depression or anhedonia lasting at least two weeks (Smith et al., 2013). In both cases and controls, we excluded individuals who had a primary diagnosis or self-reported for Bipolar

Disorder (BPD), psychosis, multiple personality disorder, autism or intellectual disability. Furthermore, we removed any controls that were taking anti-depressants, anti-psychotics or lithium.

5.2.5 Training data

Table 20 details the sample characteristics for the 80% of the data used for the discovery GWAS, and **Table 21** details the numbers of individuals for which phenotype data were available for the 15 traits on which GWAS were performed.

	Training Data
Total sample	100,478
Proportion Female	51.7%
Age	56.8 (8.01; 39 - 73)
BMI	27.4 (4.74; 13.9 - 74.7)
Neuroticism score	3.86 (3.15; 0 - 12)

Table 20. Sample characteristics for the training dataset, which represents 80% of the UK Biobank.

Phenotype	Cases	Controls	Total	% Cases
Neuroticism	-	-	82,069	-
Mood swings	41,781	56,241	98,022	42.6%
Miserableness	39,868	58,977	98,845	40.3%
Irritability	26,145	70,069	96,214	27.2%
Sensitivity	52,196	45,426	97,622	53.5%
Fed-up feelings	38,176	60,331	98,507	38.8%
Nervous feelings	20,917	77,172	98,089	21.3%
Worrier	52,932	45,006	97,938	54.0%
Tense feelings	14,976	82,587	97,563	15.4%
Worry too long	44,068	52,390	96,458	45.7%
Suffer from nerves	17,944	79,104	97,048	18.5%
Loneliness	16,367	82,626	98,993	16.5%
Guilty feelings	26,017	71,988	98,005	26.5%
Obese Yes/No	24,205	76,007	100,212	24.2%
College Yes/No	31,870	50,196	82,066	38.8%

Table 21. Number of cases and controls (total sample only for the continuous neuroticism phenotype) for the 15 phenotypes on which GWAS were performed in the training dataset (80% of the UK Biobank). Cases here refer to the number of individuals answering yes to each neuroticism endo-phenotype question, who have a BMI $\geq$ 30 kg/m^2 for the obesity phenotype, and who said they have a university degree for the college phenotype.

5.2.6 GWAS

We performed genome-wide association studies (GWAS) in the UK Biobank for the 12 neuroticism endo-phenotypes, the composite neuroticism score, obesity and whether college/university was attended. We performed linear regression for the neuroticism score GWAS and logistic regression for all other phenotypes, controlling for age, sex and the top 15 PCs in each model. GWAS were performed in PLINK-1.9 (Chang et al., 2015).

5.2.7 Polygenic risk scoring

Polygenic risk scores (PRS) were built using the PRSice software (Euesden et al., 2015) at SNP *P*-value thresholds of $0.001, 0.05, 0.1, 0.2, 0.3, 0.4$ and 0.5. Age, sex and the top 15 PCs were controlled for in the association of the scores with the phenotype in order to establish the most predictive threshold.

5.2.8 Test data PRS

Using the beta coefficients from the discovery GWAS performed in 80% of the UK Biobank ($N = 100{,}478$), we built polygenic risk scores (PRS) on the test dataset consisting of 10% of the UK Biobank ($N = 12{,}552$). For MDD and SCZ, we used the PGC MDD (Ripke et al., 2013) and PGC SCZ (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) discovery GWAS. We were not able to use the summary statistics from the latest MDD GWAS (Hyde et al., 2016) as 23andMe do not release summary statistics due to data protection. We removed all individuals in the UK Biobank who were also part of the PGC MDD discovery GWAS ($N = 8$). For the neuroticism phenotypes, as well as obesity and college, we built the PRS, and chose the most predictive score, on the phenotype itself. For example, for miserableness we chose the most predictive score of the miserableness phenotype, as opposed to of MDD, in order to obtain a score that best reflects the genetics of miserableness. For MDD and SCZ we chose the score most predictive of MDD case-control status, as SCZ PRS has previously been shown to be a better predictor of MDD than MDD PRS (Euesden et al., 2015). **Table 22** details the sample characteristics of the test dataset, and **Figure 39** details the correlations between MDD, the 12 neuroticism endo-phenotypes, obesity and college.

	Test Data
Total sample	12,552
Proportion Female	51.2%
MDD cases	955
Age	56.9 (7.96; 40 - 70)
BMI	27.3 (4.73; 15.4 - 61.9)
Neuroticism score	3.85 (3.15; 0 - 12)

Table 22. Sample characteristics of the test dataset, which represents 10% of the UK Biobank.

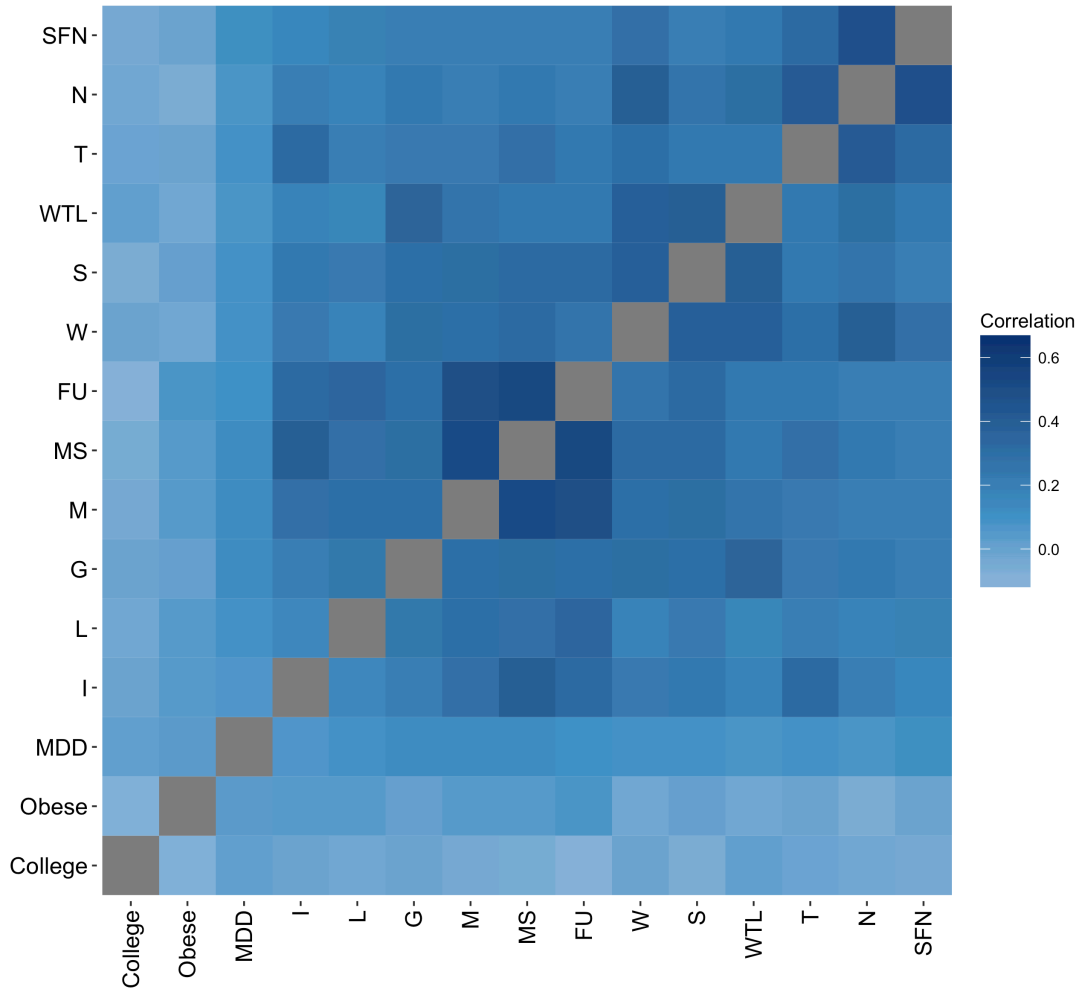

Figure 39. Correlation matrix for the 12 neuroticism endo-phenotypes, MDD case-control status, obesity and college.

5.2.9 Validation data PRS

Due to potential problems with over fitting of the predictive model(s) in the test dataset, we retained a further 10% of the UK Biobank data as a validation dataset. Again, we removed all individuals in the UK Biobank who were also part of the PGC MDD discovery GWAS ($N = 11$). We calculated the PRS for each individual on each phenotype in the validation dataset using the most predictive threshold as determined in the test data. **Table 23** details the sample characteristics of the validation dataset.

	Validation Data
Total sample	12,549
Proportion Female	51.2%
MDD cases	953
Age	56.8 (7.95; 40 - 70)
BMI	27.5 (4.82; 14.5 - 66.2)
Neuroticism score	3.84 (3.14; 0 - 12)

Table 23. Sample characteristics of the validation dataset, which represents 10% of the UK Biobank.

5.2.10 Prediction Models

We built prediction models of MDD case-control status using the PRS built in the test dataset as predictors in a logistic regression. We first fit a model with all PRS predictors (and no interaction terms), controlling for age, sex and the top 15 PCs. A stepwise forward and backward variable selection procedure was then performed to determine the most predictive subset of predictors. We use both AIC and BIC as criteria for variable selection, and the most predictive model is chosen to be the one that minimises the criterion. We compare the fit of the model against the null model containing only the control variables using a likelihood ratio test (LRT) of the nested

models. We also compare to a “MDD null model” that contains the control variables as well as the MDD PRS, and a “MDD and SCZ null model” that contains MDD and SCZ PRS as well as the control variables. We compare the Nagelkerke’s psuedo-R^2 between the fitted and null models. We also consider two-way interaction terms between the PRS predictors, as well as allowing for interaction between them and the age and sex covariates. As with the main effects models, we use stepwise variable selection procedures to determine the most predictive models and compare to the null models using likelihood ratio tests (LRTs) and Nagelkerke’s pseudo-R^2.

In the validation data, we rebuild the models chosen to be the most predictive of MDD case-control status from the stepwise variable selection procedures in order to determine whether the models replicate in an independent dataset.

We repeat the above procedure using phenotypic data as predictors, as opposed to the PRS predictors, in this case controlling for age and sex in our fitted models. We also consider a mixture of genetic (PRS) and phenotypic predictors, where the top 15 PCs are controlled for in addition to age and sex.

Figure 40 provides an illustrative overview of the analyses to be performed in the UK Biobank.

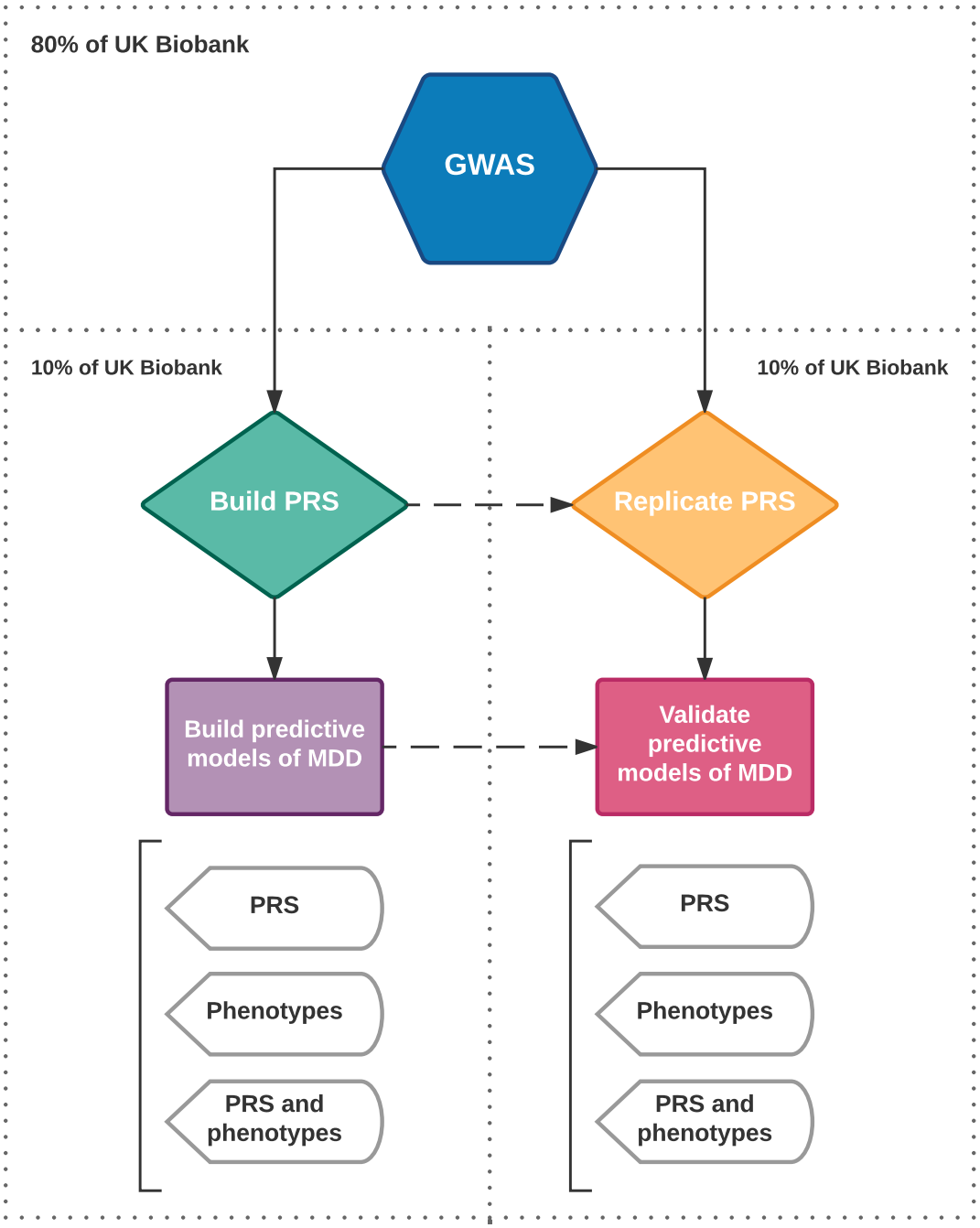

Figure 40. Illustrative overview of the analyses to be performed.

5.3 Results

5.3.1 GWAS

GWAS were performed in the training dataset for the 12 neuroticism endo-phenotypes, the neuroticism score (sum across all 12 questions), obesity and college, in order to obtain the beta coefficients from which to build the PRS in the test dataset. **Table 24** details the number of independent genome-wide significant hits identified by each of the GWAS. Independence here is defined as there being no other genome-wide significant SNP within a 500kb window centered on that SNP. **Figure 41** shows the Manhattan plots for the GWAS on neuroticism, worrier (W), obesity and college, which identified the largest number of genome-wide significant findings.

Phenotype	Sample Size	GWAS Hits
Neuroticism	82,069	11
Mood swings	98,022	2
Miserableness	98,845	3
Irritability	96,214	0
Sensitivity	97,622	0
Fed-up feelings	98,507	0
Nervous feelings	98,089	0
Worrier	97,938	10
Tense feelings	97,563	0
Worry too long	96,458	1
Suffer from nerves	97,048	0
Loneliness	98,993	0
Guilty feelings	98,005	1
Obese Yes/No	100,212	12
College Yes/No	82,066	5

Table 24. Number of independent genome-wide significant associations identified by each of the GWAS performed in the training dataset (80% of the UK Biobank).

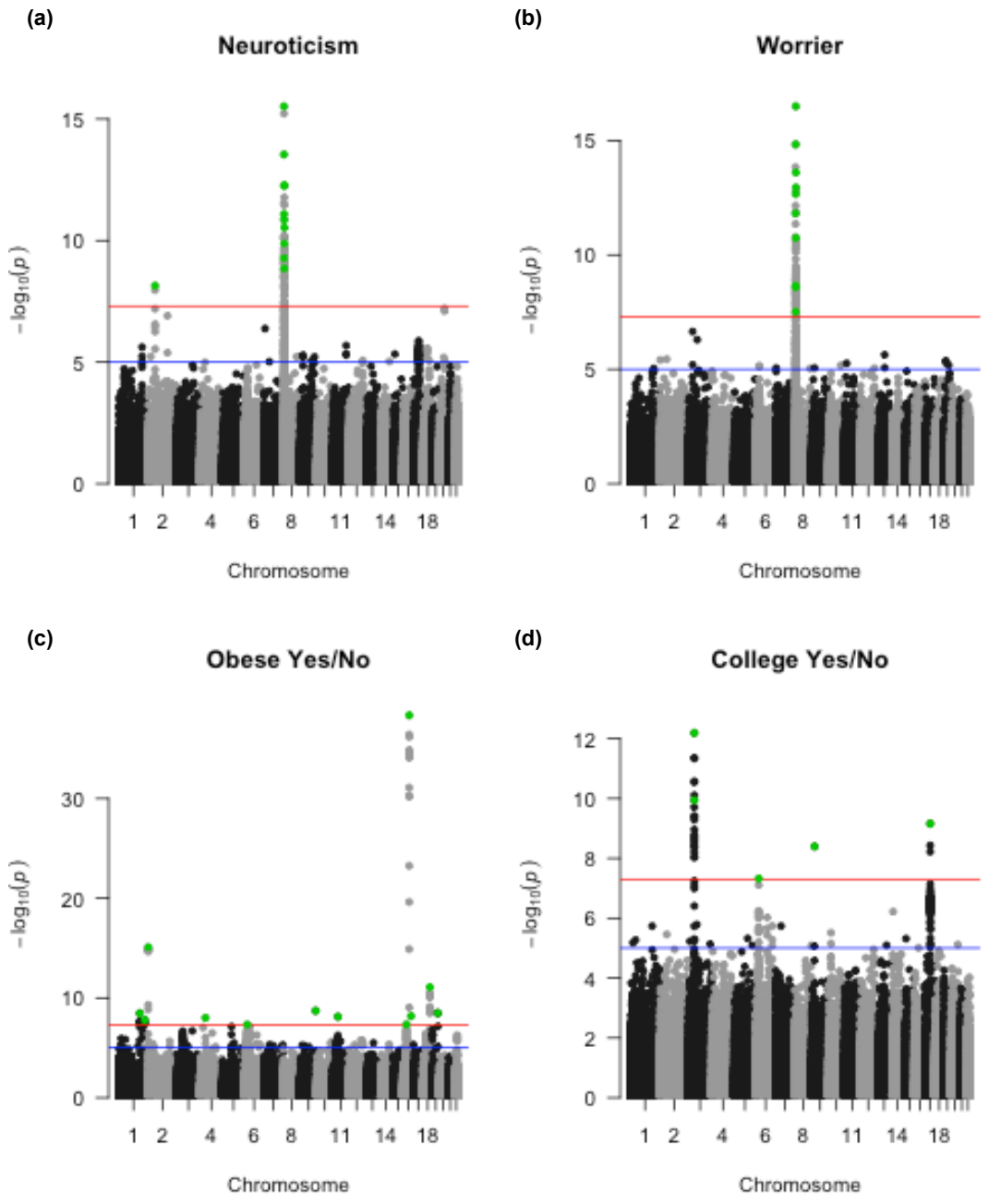

Figure 41. Manhattan plots for the GWAS on neuroticism, worrier (W), obesity and college, performed in the training dataset (80% of the UK Biobank). The green points represent the independent, genome-wide significant signals.

Table 25 contains the genome-wide significant associations for the neuroticism GWAS, and **Table 26** details the genome-wide significant associations for the worrier (W) GWAS. We observe that the same signal on chromosome 8 is present in both GWAS. This top hit was also identified in a recent neuroticism meta-analysis (Smith

et al., 2016). The signal on chromosome 2 for neuroticism is not observed in the worrier (W) GWAS.

SNP	CHR	Position (Mb)	*P*-value
rs12547493	8	8.66	3.01×10^{-16}
rs7826660	8	8.38	2.87×10^{-14}
rs12679529	8	10.8	5.25×10^{-13}
rs3808509	8	11.2	5.61×10^{-13}
rs2929453	8	9.08	8.17×10^{-12}
rs9969657	8	10.5	1.40×10^{-11}
rs11998678	8	11.8	2.89×10^{-11}
rs13280813	8	11.4	1.32×10^{-10}
rs2980437	8	8.09	5.31×10^{-10}
rs3088186	8	10.2	1.40×10^{-9}
rs2678890	2	58.2	7.21×10^{-9}

Table 25. Genome-wide significant, independent hits for the neuroticism GWAS performed in the training dataset (80% of the UK Biobank).

SNP	CHR	Position (Mb)	*P*-value
rs12679529	8	10.8	3.19×10^{-17}
rs12547493	8	8.66	1.49×10^{-15}
rs3808509	8	11.2	2.52×10^{-14}
rs11998678	8	11.8	1.15×10^{-13}
rs9969657	8	10.5	2.18×10^{-13}
rs2976940	8	8.28	1.52×10^{-12}
rs13280813	8	11.4	1.82×10^{-11}
rs28649568	8	8.96	2.23×10^{-9}
rs3088186	8	10.2	2.56×10^{-9}
rs656319	8	9.81	3.06×10^{-8}

Table 26. Genome-wide significant, independent hits for the worrier (W) GWAS performed in the training dataset (80% of the UK Biobank).

Table 27 details the genome-wide significant SNPs for the obesity GWAS, and **Table 28** details the genome-wide significant SNPs for the college GWAS.

SNP	CHR	Position (Mb)	*P*-value
rs1421085	16	53.8	4.57×10^{-39}
rs2867125	2	0.62	8.11×10^{-16}
rs10871777	18	57.9	8.28×10^{-12}
rs12343000	9	131.0	1.92×10^{-9}
rs34783010	19	46.2	3.27×10^{-9}
rs543874	1	177.9	3.43×10^{-9}
rs1364063	16	69.6	6.55×10^{-9}
rs1061810	11	43.9	7.55×10^{-9}
rs10938397	4	45.2	9.70×10^{-9}
rs2061154	1	219.7	1.60×10^{-8}
rs4788190	16	29.9	4.63×10^{-8}
rs10484439	6	26.3	4.75×10^{-8}

Table 27. Genome-wide significant, independent hits for the obesity GWAS performed in the training dataset (80% of the UK Biobank).

SNP	CHR	Position (Mb)	*P*-value
rs868891	3	49.9	6.46×10^{-13}
rs4625	3	49.6	1.14×10^{-10}
rs55663797	17	43.5	6.93×10^{-10}
rs12554512	9	23.4	4.02×10^{-9}
rs806794	6	26.2	4.77×10^{-8}

Table 28. Genome-wide significant, independent hits for the college GWAS performed in the training dataset (80% of the UK Biobank).

5.3.2 Test data PRS

Polygenic risk scores (PRS) were built in the test dataset using the SNP coefficients determined by the GWAS performed in the training data. MDD and SCZ discovery GWAS from the Psychiatric Genomics Consortium (PGC) were also used (Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). **Table 29** details the most predictive thresholds for each phenotype, as well as the phenotypic variance explained by this most predictive score and the corresponding *P*-value of association.

Phenotype	Sample Size	Threshold	*P*-value	R^2	No. SNPs
Neuroticism	10,219	0.05	4.37×10^{-18}	0.0071	18,721
Mood swings	12,241	0.3	3.13×10^{-7}	0.0028	82,856
Miserableness	12,333	0.05	7.39×10^{-7}	0.0026	18,426
Irritability	12,006	0.3	5.77×10^{-10}	0.0046	82,212
Sensitivity	12,195	0.4	8.54×10^{-12}	0.0049	103,413
Fed-up feelings	12,297	0.2	1.03×10^{-13}	0.0060	59,799
Nervous feelings	12,227	0.5	6.64×10^{-12}	0.0060	121,426
Worrier	12,217	0.3	4.08×10^{-12}	0.0051	82,249
Tense feelings	12,195	0.5	1.94×10^{-9}	0.0051	121,374
Worry too long	12,050	0.5	1.22×10^{-13}	0.0060	121,754
Suffer from nerves	12,105	0.05	1.15×10^{-4}	0.0020	17,242
Loneliness	12,369	0.4	3.60×10^{-8}	0.0041	102,784
Guilty feelings	12,220	0.1	1.91×10^{-5}	0.0021	32,499
Obese Yes/No	12,521	0.2	1.20×10^{-48}	0.0260	62,223
College Yes/No	10,261	0.4	6.33×10^{-37}	0.0213	104,799
MDD	12,552	0.001	2.71×10^{-1}	0.0002	134
SCZ	12,552	0.3	1.94×10^{-2}	0.0010	81,326

Table 29. Most predictive polygenic risk score (PRS) thresholds for each phenotype, and the corresponding variance explained (R^2) and *P*-value. For the neuroticism phenotypes, obesity and college, the threshold was chosen so that the PRS was most predictive of the same phenotype. For MDD and SCZ the threshold was chosen for the most predictive PRS of MDD. Tested thresholds were 0.001, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5.

As shown in **Table 29**, the best threshold PRS for obese (yes/no) and college (yes/no) have the largest variance explained out of all traits considered here at 2.6% ($P = 1.20 \times 10^{-48}$) and 2.13% ($P = 6.33 \times 10^{-37}$) respectively. Interestingly the schizophrenia (SCZ) PRS, which has been shown to have a variance explained of around 2.1% ($P = 2.10 \times 10^{-12}$) for predicting MDD status in one sample (Euesden et al., 2015), only has an R^2 of 0.1% ($P = 1.94 \times 10^{-2}$) in this sample. This is likely due to the low prevalence of SCZ in the UK Biobank cohort. While the neuroticism PRS only has a modest R^2 of 0.71% ($P = 4.37 \times 10^{-18}$), this is in line with previous studies (Smith et al., 2016). Of the component measures of neuroticism, fed-up feelings, nervous feelings and worry too long have the best predictive power (R^2 = 0.6%; $P = 1.03 \times 10^{-13}$, $P = 6.64 \times 10^{-12}$ and $P = 1.22 \times 10^{-13}$ respectively), while mood swings and miserableness have the lowest variance explained (R^2 of 0.28% ($P = 3.13 \times 10^{-7}$) and 0.26% ($P = 7.39 \times 10^{-7}$) respectively). The highest variance explained that has been achieved by a behavioral trait PRS to date is 9% for educational achievement (Selzam et al., 2016), which puts the modest prediction achieved for the neuroticism traits into context. However, educational achievement has been estimated to have a twin heritability of around 60% (Krapohl et al., 2014), suggesting a large genetic component that may not be replicated in other traits. The PRS with the smallest R^2 in this sample is MDD at 0.02% ($P = 0.271$). In a recent mega-analysis, the MDD PRS has been shown to explain 0.6% of variance in MDD case/control status ($P < 10^{-6}$) (Ripke et al., 2013), suggesting that the MDD PRS could be underpowered in this sample due to the heterogeneous nature of the MDD phenotype.

The predictive power of the PRS in this sample are modest, though the relative performance of the MDD-correlated-trait PRS compared to that of MDD could suggest that these PRS may provide more predictive power than MDD itself. In particular, using a combination of the PRS in a multivariate model could leverage the

individually limited variance explained to identify an overall more predictive model of MDD. However, the low predictive power that we have observed here clearly indicates that these PRS have almost no direct clinical benefit as yet, but should instead be viewed as potentially providing insights into disease and trait aetiology, which may subsequently have clinical use.

Figure 42 shows the PRSice (Euesden et al., 2015) bar plots corresponding to PRS on MDD (**Figure 42a**) and SCZ (**Figure 42b**) built in the test dataset, and predicting MDD case/control status.

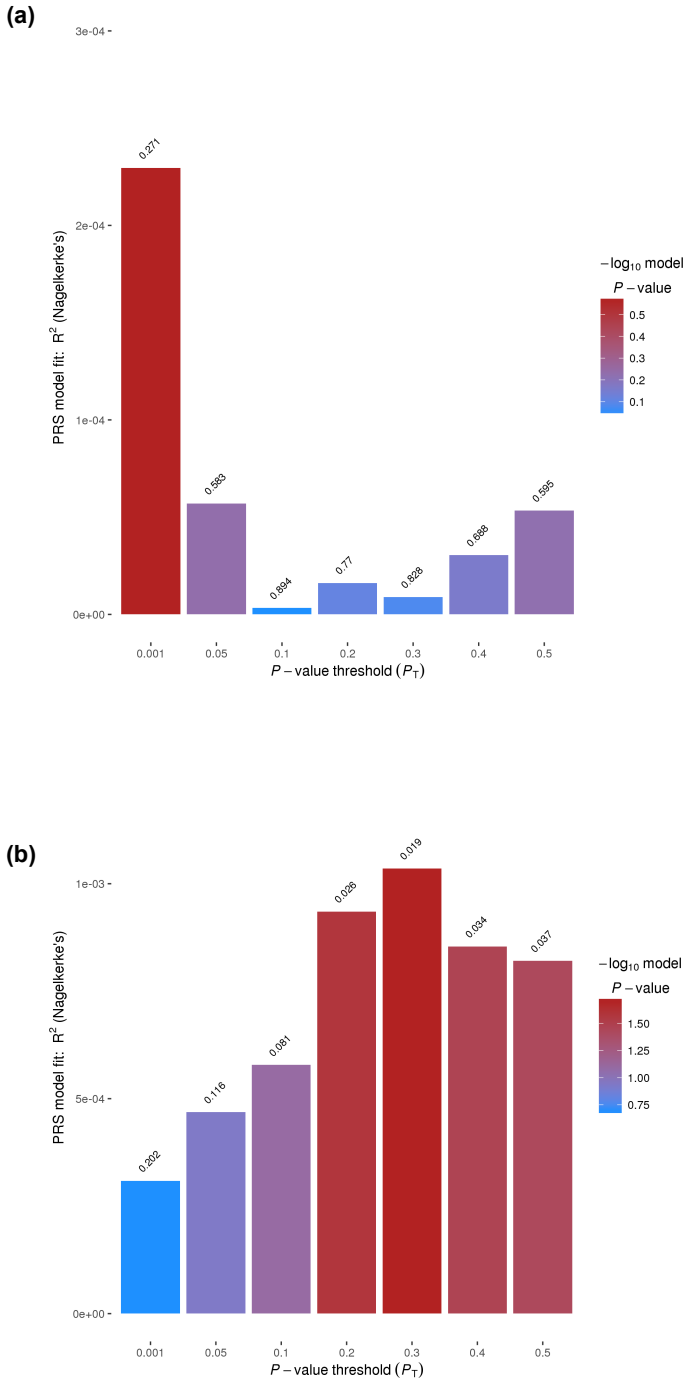

Figure 42. PRSice bar plots for the PRS built in the test dataset for: **(a)** MDD and **(b)** SCZ, illustrating the different SNP *P*-value thresholds and their prediction of MDD case-control status.

Figure 43 shows the PRSice (Euesden et al., 2015) bar plots for the most predictive PRS thresholds across all phenotypes on which PRS were built.

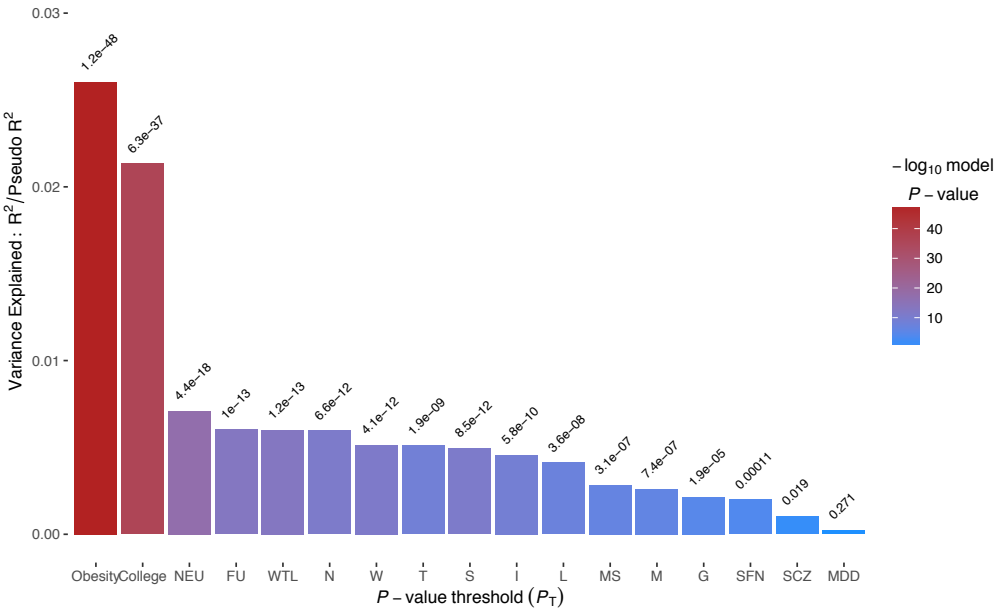

Figure 43. PRSice bar plots for the most predictive PRS as determined in the test dataset, across all phenotypes on which PRS were built.

5.3.3 Prediction models of MDD using multiple PRS predictors

Previous analyses have shown that the PRS for MDD has limited predictive power of MDD case-control status (Ripke et al., 2013). Here we aim to build a more powerful predictive model of MDD using both MDD and SCZ PRS, and PRS for other correlated traits. We use variable selection techniques to find the most predictive models, and consider both the PRS main effects and two-way interactions between PRS. We first build a predictive model in the test dataset, with an independent dataset reserved for cross-validation.

5.3.3.1 PRS: Main effect predictors

We fit a logistic regression model with all PRS included as predictors, and include age, sex and the top 15 PCs as control variables:

$$MDD_{status} \sim MDD_{PRS} + SCZ_{PRS} + NEU_{PRS} + MS_{PRS} + M_{PRS} + I_{PRS} + S_{PRS} + FU_{PRS} + N_{PRS} + W_{PRS} + T_{PRS} + WTL_{PRS} + SFN_{PRS} + L_{PRS} + G_{PRS} + Obese_{PRS} + College_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

Using the forward and backward stepwise variable selection procedure, the most predictive model of MDD using Akaike Information Criterion (AIC) to assess model fit was:

$$MDD_{status} \sim MDD_{PRS} + SCZ_{PRS} + N_{PRS} + W_{PRS} + T_{PRS} + WTL_{PRS} + L_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

We compared the fit of this model to that of the null model, containing only age, sex and the top 15 PCs as covariates. We also compared to the MDD null model, which in addition includes the MDD PRS as a predictor, and to the MDD and SCZ null model, which includes the MDD and SCZ PRS as predictors in addition to the control variables. We compare to the MDD null and MDD/SCZ null in order to determine whether there is any gain over the standard approach of using the MDD PRS, or indeed the MDD and SCZ PRS, in predicting MDD case/control status. The corresponding Nagelkerke's pseudo-R^2 and likelihood ratio test *P*-values are given in **Table 30**.

Model	**R^2**	***P*-value**
Main effects (AIC) vs. Null model	0.00877	2.90×10^{-5}
Main effects (AIC) vs. MDD null model	0.00776	6.06×10^{-5}
Main effects (AIC) vs. MDD and SCZ null model	0.00697	8.67×10^{-5}

Table 30. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) main effect model compared to the null model, the MDD null model and the MDD and SCZ null model.

We also performed the same procedure but instead using the Bayesian Information Criterion (BIC) to assess model fit. BIC is similar to AIC, but penalises the inclusion of additional predictors in the model more severely. The resulting model from the stepwise procedure using BIC is:

$$MDD_{status} \sim L_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

L here refers to the loneliness neuroticism endo-phenotype. As the MDD and SCZ PRS were not retained in the model after the variable selection procedure, we compared the fit of this model only to that of the null model, the results of which are given in **Table 31**.

Model	**R^2**	***P*-value**
Main effects (BIC) vs. Null model	0.00252	2.14×10^{-3}
Main effects (BIC) vs. MDD null model	-	-
Main effects (BIC) vs. MDD and SCZ null model	-	-

Table 31. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) main effect model compared to the null model. The MDD and SCZ PRS were not retained in the model, so comparisons with the MDD null model and the MDD and SCZ null model were not possible.

5.3.3.2 PRS: Two-way interaction models

We next fit a logistic regression model with all the same predictors as in the previous main effects model, but we also include two-way interactions between the PRS predictors, as well as interactions between them and the age and sex covariates. There are a total of 171 two-way interaction terms in the model in addition to the main effects terms, on which the variable selection procedure is performed.

The stepwise variable selection two-way interaction best-fit model as determined by AIC is (PRS subscript omitted here for ease of reading):

$$MDD_{status} \sim MS + M + S + FU + N + W + T + WTL + SFN + L + G + NEU + MDD + SCZ + Obese + College + Age{:}M + Age{:}FU + Age{:}W + Age{:}SCZ + Age{:}College + Sex{:}M + Sex{:}FU + Sex{:}T + Sex{:}SFN + MS{:}M + MS{:}FU + MS{:}WTL + M{:}L + S{:}WTL + S{:}G + S{:}NEU + S{:}SCZ + S{:}Obese + FU{:}N + FU{:}W + FU{:}T + FU{:}WTL + FU{:}College + N{:}G + N{:}SCZ + N{:}Obese + W{:}L + W{:}G + W{:}NEU + W{:}SCZ + T{:}WTL + T{:}L + T{:}NEU + SFN{:}L + SFN{:}College + L{:}SCZ + NEU{:}College + MDD{:}Obese + Age + Sex + PC_1 + \cdots + PC_{15}$$

We compare this two-way interaction model to the three null models as defined previously. The results of the likelihood ratio tests (LRTs) and comparison of Nagelkerke's pseudo-R^2 between these models are given in **Table 32**.

Model	**R^2**	***P*-value**
Interaction model (AIC) vs. Null model	0.0414	8.51×10^{-12}
Interaction model (AIC) vs. MDD null model	0.0403	1.75×10^{-11}
Interaction model (AIC) vs. MDD and SCZ null model	0.0396	2.73×10^{-11}

Table 32. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) two-way interaction model compared to the null model, the MDD null model and the MDD and SCZ null model.

The stepwise variable selection two-way interaction model as determined by BIC is:

$$MDD_{status} \sim FU + Age{:}FU + Age + Sex + PC_1 + \cdots + PC_{15}$$

FU here refers to the fed-up neuroticism endo-phenotype. The results of the likelihood ratio tests (LRTs) and comparison of Nagelkerke's pseudo-R^2 between this model and the null models are given in **Table 33**.

Model	R^2	*P*-value
Interaction (BIC) vs. Null model	0.00349	1.48×10^{-3}
Interaction (BIC) vs. MDD null model	-	-
Interaction (BIC) vs. MDD and SCZ null model	-	-

Table 33. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) two-way interaction model compared to the null model. The MDD and SCZ PRS were not retained in the model, so comparisons with the MDD null model and the MDD and SCZ null model were not possible.

5.3.4 Validation of PRS prediction models

PRS were rebuilt in the validation dataset using the most predictive thresholds as determined in the test data. The SNP coefficients were obtained in the discovery GWAS performed in the training data and from the PGC for MDD and SCZ (Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). **Figure 44** shows the PRSice (Euesden et al., 2015) bar plots for the most predictive thresholds across all phenotypes for which PRS were built in the validation dataset. Further details for the PRS are provided in **Table 34**.

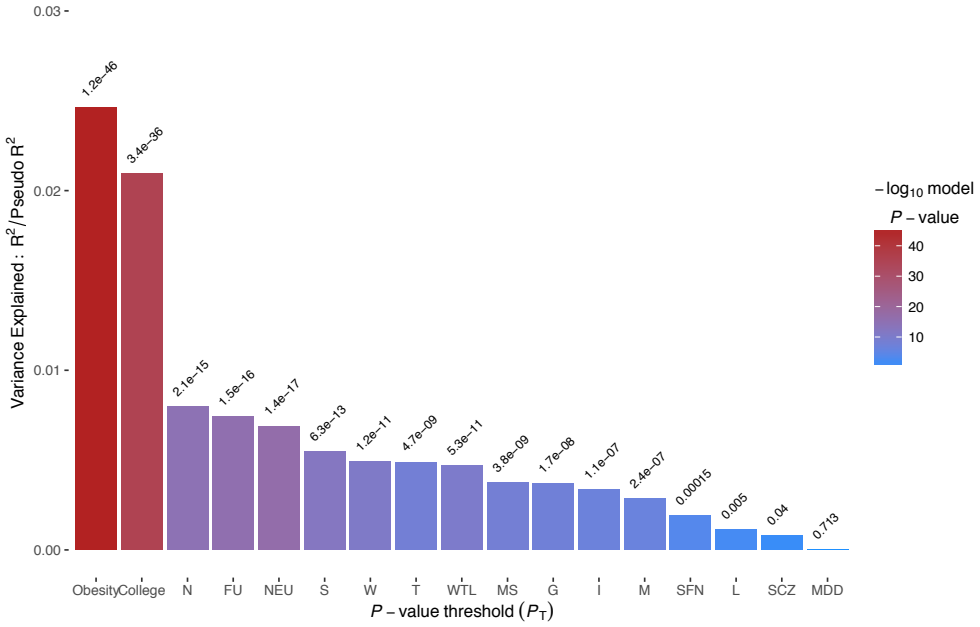

Figure 44. PRSice bar plots for the most predictive PRS as determined in the test dataset and built in the validation dataset.

Phenotype	Sample Size	Threshold	*P*-value	R^2	No. SNPs
Neuroticism	10,203	0.05	1.44×10^{-17}	0.00690	18,721
Mood swings	12,264	0.3	3.81×10^{-9}	0.00376	82,856
Miserableness	12,345	0.05	2.37×10^{-7}	0.00284	18,426
Irritability	11,997	0.3	1.09×10^{-7}	0.00335	82,212
Sensitivity	12,187	0.4	6.31×10^{-13}	0.00547	103,413
Fed-up feelings	12,309	0.2	1.53×10^{-16}	0.00743	59,799
Nervous feelings	12,235	0.5	2.09×10^{-15}	0.00796	121,426
Worrier	12,233	0.3	1.17×10^{-11}	0.00491	82,249
Tense feelings	12,162	0.5	4.74×10^{-9}	0.00489	121,374
Worry too long	12,032	0.5	5.29×10^{-11}	0.00469	121,754
Suffer from nerves	12,078	0.05	1.49×10^{-4}	0.00193	17,242
Loneliness	12,363	0.4	4.51×10^{-3}	0.00112	102,784
Guilty feelings	12,212	0.1	1.69×10^{-8}	0.00370	32,499
Obese Yes/No	12,520	0.2	1.21×10^{-46}	0.02463	62,223
College Yes/No	10,215	0.4	3.40×10^{-36}	0.02094	104,799
MDD	12,549	0.001	7.13×10^{-1}	0.00003	134
SCZ	12,549	0.3	4.02×10^{-2}	0.00080	81,326

Table 34. Most predictive polygenic risk score (PRS) thresholds for each phenotype as determined in the test dataset. Scores were then rebuilt at these thresholds in the validation dataset, and the corresponding variance explained (R^2) and *P*-value for each phenotype are given here.

The models that were determined to be most predictive of MDD case-control status in the test dataset by the stepwise variable selection procedures were then applied to the validation dataset using the PRS as detailed in **Table 34**.

5.3.4.1 PRS: Main effect predictors

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (AIC) main effects model as fitted in the validation dataset are given in **Table 35**.

Model	R^2	*P*-value
Main effects (AIC) vs. Null model	0.00154	0.567
Main effects (AIC) vs. MDD null model	0.00154	0.450
Main effects (AIC) vs. MDD and SCZ null model	0.00076	0.723

Table 35. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) main effect model as determined in the test dataset and fitted in the validation dataset, compared to the null model, the MDD null model and the MDD and SCZ null model.

The Nagelkerke's pseudo-R^2 and liikehood ratio test (LRT) *P*-values from the comparison of the stepwise (BIC) main effect model as fitted in the validation dataset are given in **Table 36**.

Model	R^2	*P*-value
Main effects (BIC) vs. Null model	0.000104	0.532
Main effects (BIC) vs. MDD null model	-	-
Main effect (BIC) vs. MDD and SCZ null model	-	-

Table 36. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) main effect model as determined in the test dataset and fitted in the validation dataset, compared to the null model.

We observe that neither the AIC or BIC selected main effect models show significant prediction in the validation dataset compared to the null model containing only the control variables. In addition, the AIC main effect model is not significant when compared to the MDD null model and the MDD and SCZ null model.

5.3.4.2 PRS: Two-way interactions

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (AIC) two-way interaction model as fitted in the validation dataset are given in **Table 37**.

Model	R^2	*P*-value
Interaction model (AIC) vs. Null model	0.0125	0.740
Interaction model (AIC) vs. MDD null model	0.0125	0.707
Interaction model (AIC) vs. MDD and SCZ null model	0.0118	0.775

Table 37. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) two-way interaction model as determined in the test dataset and fitted in the validation dataset, compared to the null model, the MDD null model and the MDD and SCZ null model.

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (BIC) two-way interaction model as fitted in the validation dataset are given in **Table 38**.

Model	R^2	*P*-value
Interaction model (BIC) vs. Null model	0.00114	0.119
Interaction model (BIC) vs. MDD null model	-	-
Interaction model (BIC) vs. MDD and SCZ null model	-	-

Table 38. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) two-way interaction model as determined in the test dataset and fitted in the validation dataset, compared to the null model.

Again, we observe that neither the AIC nor BIC selected two-way interaction models replicate significantly in the validation dataset when compared to the null model. Likewise, for the AIC two-way interaction model compared to the MDD null model and the MDD and SCZ null model.

5.3.5 Prediction models of MDD using phenotype-only data

We now repeat the same procedure as for the PRS analyses above but using phenotype predictors instead, to test whether we can obtain better prediction with phenotype data alone.

5.3.5.1 Neuroticism multi-trait analyses

We first compare the use of the continuous neuroticism score (NEU) to that of using the 12 neuroticism endo-phenotypes in the prediction of MDD. Complete phenotype data across all 12 neuroticism endo-phenotypes were available on 10,219 individuals. We fit a logistic regression model of NEU predicting MDD case-control status, controlling for age and sex:

$$MDD_{status} \sim NEU + Age + Sex$$

We compare this model to the null model containing only the age and sex covariates in a likelihood ratio test (LRT), as well as comparing the Nagelkerke's pseudo-R^2 (see **Table 39**).

Model	**R^2**	***P*-value**
NEU model vs. Null model	0.0561	1.33×10^{-56}

Table 39. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the neuroticism score model compared to the null model.

Next, we fit a logistic regression model of the 12 neuroticism endo-phenotypes predicting MDD case/control status, controlling for age and sex:

$$MDD_{status} \sim MS + M + I + S + FU + N + W + T + WTL + SFN + L + G + Age + Sex$$

Table 40 details the comparison of this model with the null model in a likelihood ratio test (LRT), as well as the comparison of Nagelkerke's pseudo-R^2.

Model	R^2	*P*-value
NEU component model vs. Null model	0.0650	2.23×10^{-55}

Table 40. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the neuroticism endo-phenotype model compared to the null model.

We also performed stepwise variable selection procedures on this model using both AIC and BIC as inclusion criterion. The resulting model from the AIC procedure is:

$$MDD_{status} \sim MS + M + W + T + SFN + L + G + Age + Sex$$

The likelihood ratio test (LRT) *P*-value of the comparison of this model with the null model is given in **Table 41**, along with Nagelkerke's pseudo-R^2 for this model.

Model	R^2	*P*-value
NEU component model (AIC) vs. Null model	0.0644	1.26×10^{-58}

Table 41. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype model compared to the null model.

The best-fitting model from the BIC variable selection procedure is:

$$MDD_{status} \sim MS + M + SFN + G + Age + Sex$$

Table 42 details the likelihood ratio test (LRT) of the comparison of this model with the null model along with Nagelkerke's pseudo-R^2.

Model	R^2	*P*-value
NEU component model (BIC) vs. Null model	0.0613	2.43×10^{-58}

Table 42. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype model compared to the null model.

We observe that for the model with all 12 neuroticism endo-phenotypes as predictors, and for both the models containing a subset of these predictors as determined by AIC and BIC variable selection criterion, the prediction of MDD case/control status is greater than that of using the continuous neuroticism score alone (these models all achieve a larger phenotypic variance explained). This suggests that taking a multivariate approach may yield better prediction, and that using stepwise variable selection procedures can inform us as to the subtypes of phenotypes that most correlate with the disease outcome and that should not be merely over-fit.

5.3.5.2 Prediction of MDD using phenotype data

Here we build predictive models of MDD case-control status using phenotype data alone. The phenotypes we consider are the 12 neuroticism endo-phenotypes, obesity and whether college was attended, and age and sex are used as control variables. Complete phenotype data were available for $8{,}459$ individuals.

5.3.5.3 Main effects models

We perform the same variable selection procedures as with the PRS predictors but instead using the phenotype data to predict MDD case-control status. First we fit a logistic regression model with all phenotypes as predictors, controlling for age and sex:

$$MDD_{status} \sim MS + M + I + S + FU + N + W + T + WTL + SFN + L + G + Obese + College + Age + Sex$$

A forward and backward variable selection procedure was applied to this full model using both AIC and BIC criterion. The model selected based on AIC is:

$$MDD_{status} \sim MS + M + N + W + T + SFN + L + G + Obese + College + Age + Sex$$

We compare the fit of this model to that of the null model containing only the age and sex covariates in a likelihood ratio test (LRT). **Table 43** details the results of the LRT as well as Nagelkerke's pseudo-R^2 for this model.

Model	**R^2**	***P*-value**
Main effects (AIC) vs. Null model	0.0608	2.19×10^{-43}

Table 43. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) endophenotype, obesity and college model compared to the null model.

The best-fitting model selected based on BIC is:

$$MDD_{status} \sim MS + SFN + G + Age + Sex$$

Table 44 details the comparison of the fit of this model to that of the null model in a likelihood ratio test (LRT), as well as Nagelkerke's pseudo-R^2 for this model.

Model	**R^2**	***P*-value**
Main effects model vs. Null model	0.0523	3.12×10^{-42}

Table 44. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) endophenotype, obesity and college model compared to the null model.

5.3.5.4 Two-way interactions

We next fit a logistic regression model with all the same predictors as in the previous main effects model, but we also include two-way interactions between the phenotypes, as well as with the age and sex covariates. There are a total of 120 two-way interaction terms in the model in addition to the 16 main effect terms and the control variables, on which the variable selection procedure is performed.

The stepwise variable selection two-way interaction model as determined by AIC is:

$$MDD_{status} \sim MS + M + I + S + FU + N + W + T + WTL + SFN + L + G + Obese + College + MS:N + MS:G + MS:Age + M:N + M:W + M:T + M:L + M:Sex + I:SFN + I:L + S:Age + FU:SFN + FU:College + N:W + T:SFN + WTL:L + WTL:Sex + WTL:College + L:G + L:Age + G:Sex + G:College + Age:College + Sex:Obese + Age + Sex$$

We compare the fit of this model to that of the null model containing only the age and sex covariates using a likelihood ratio test (LRT) and compare Nagelkerke's pseudo-R^2 (see **Table 45**), as well as comparing with the full main effect model (all phenotypes included as predictors), and the AIC selected main-effect model.

Model	**R^2**	***P*-value**
Interaction model (AIC) vs. Null model	0.0902	2.39×10^{-50}
Interaction model (AIC) vs. Full model	0.0289	4.46×10^{-13}
Interaction model (AIC) vs. Main effects (AIC) model	0.0295	3.78×10^{-12}

Table 45. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, obesity and college two-way interaction model compared to the null model, the full model and the stepwise (AIC) model.

The stepwise variable selection two-way interaction model as determined by BIC is:

$$MDD_{status} \sim MS + M + FU + N + WTL + SFN + G + College + M:N + FU:College + WTL:College + Age + Sex$$

Table 46 details the comparison of this model to that of the null model containing only the age and sex covariates using a likelihood ratio test (LRT) and Nagelkerke's

pseudo-R^2, as well as a comparison with the BIC selected main effect model. We did not compare to the full model here, as not all predictors were retained in the model.

Model	R^2	*P*-value
Interaction model (BIC) vs. Null model	0.0639	3.79×10^{-45}
Interaction model (BIC) vs. Main effects (BIC) model	0.0115	6.14×10^{-7}

Table 46. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, obesity and college two-way interaction model compared to the null model and the stepwise (BIC) model.

5.3.6 Validation of phenotype models

As with the PRS predictive models, we now aim to validate the models built in the test dataset using the phenotypic data only.

5.3.6.1 Neuroticism multi-trait models

We perform the same comparison models of the total neuroticism score (NEU) predicting MDD and the 12 neuroticism endo-phenotypes predicting MDD in the validation dataset.

The likelihood ratio test (LRT) of the NEU score model compared to the null model containing only age and sex as predictors, as well as Nagelkerke's pseudo-R^2 for this model, are given in **Table 47**.

Model	R^2	*P*-value
NEU model vs. Null model	0.0484	4.55×10^{-48}

Table 47. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the neuroticism score model compared to the null model in the validation dataset.

The likelihood ratio test (LRT) for the model with the 12 neuroticism endo-phenotypes as predictors compared to the null model, as well as Nagelkerke's pseudo-R^2 for this model, are given in **Table 48**.

Model	R^2	*P*-value
NEU component model vs. Null model	0.0580	1.35×10^{-47}

Table 48. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the neuroticism endo-phenotype model compared to the null model in the validation dataset.

Table 49 details the likelihood ratio test (LRT) results and the Nagelkerke's pseudo-R^2 for the AIC selected neuroticism endo-phenotype model as fitted in the validation dataset.

Model	R^2	*P*-value
NEU component model (AIC) vs. Null model	0.0572	1.54×10^{-50}

Table 49. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the AIC selected neuroticism endo-phenotype model compared to the null model in the validation dataset.

Table 50 details the likelihood ratio test (LRT) results and the Nagelkerke's pseudo-R^2 for the BIC selected neuroticism endo-phenotype model as fitted in the validation dataset.

Model	R^2	*P*-value
NEU component model (BIC) vs. Null model	0.0515	1.07×10^{-47}

Table 50. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the BIC selected neuroticism endo-phenotype model compared to the null model in the validation dataset.

We see that in all cases the models replicate in the independent validation dataset, and that the models built using the 12 neuroticism endo-phenotypes (and subsets) provide better prediction of MDD case/control status than using the total neuroticism score.

5.3.6.2 Main effect models

We now rebuild the phenotype predictor models of MDD case/control status in the validation dataset.

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (AIC) main effect model with the null model as fitted in the validation dataset are given in **Table 51**.

Model	**R^2**	***P*-value**
Main effects model (AIC) vs. Null model	0.0646	8.65×10^{-47}

Table 51. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, obesity and college model compared to the null model in the validation dataset.

Similarly, the results of the likelihood ratio test (LRT) of the BIC selected main effect model from the test dataset as fitted in the validation dataset are given in **Table 52**.

Model	**R^2**	***P*-value**
Main effects model (BIC) vs. Null model	0.0450	1.88×10^{-36}

Table 52. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, obesity and college model compared to the null model in the validation dataset.

We observe that for both the AIC and BIC selected main effect models, the models are significantly replicated in the independent validation dataset, and from these results the AIC penalty appears to be most appropriate because the model from the AIC procedure is more predictive than that from the BIC.

5.3.6.3 Two-way interactions

Next we fit the AIC selected two-way interaction model in the independent validation dataset, and compare this model to the null model containing only age and sex

covariates, as well as the full model that includes all phenotype predictors and the stepwise main effects model, which was the result of the AIC variable selection procedure in the test dataset. The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from these comparisons are given in **Table 53**.

Model	**R^2**	***P*-value**
Interaction model (AIC) vs. Null model	0.0728	8.69×10^{-38}
Interaction model (AIC) vs. Full model	0.00695	0.317
Interaction model (AIC) vs. Main effects (AIC) model	0.00812	0.309

Table 53. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, obesity and college two-way interaction model compared to the null model, the full model and the stepwise (AIC) model in the validation dataset.

We perform the same validation for the BIC selected two-way interaction model, this time comparing only to the null model and the BIC selected stepwise main effect model (see **Table 54**). We do not compare to the full model here as not all phenotype predictors were retained in the model.

Model	**R^2**	***P*-value**
Interaction model (BIC) vs. Null model	0.0595	5.18×10^{-42}
Interaction model (BIC) vs. Stepwise model	0.0146	3.42×10^{-9}

Table 54. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, obesity and college two-way interaction model compared to the null model and the stepwise (BIC) model in the validation dataset.

We observe that, while the AIC selected two-way interaction model offers significantly greater prediction when compared to the null model in the validation dataset, the prediction is not significantly greater than the full main effects model (containing all phenotype predictors), or the AIC selected main effect model. This suggests that the model chosen by the AIC variable selection procedure in the test dataset was over fit, and thus does not replicate in an independent dataset.

For the BIC selected two-way interaction model, we see that for both the comparison with the null model and with the BIC selected main effect model, the BIC selected two-way interaction model offers better prediction of MDD case-control status. Using BIC as selection criteria is a more stringent approach, and will most often result in fewer predictors being retained in the model. This reduces the chance of over fitting, thus maximising the chance of replication in an independent dataset. However, when fitting main effects models we observed that using AIC lead to a more predictive model than the BIC interaction model. These results suggest that the AIC selected main effect model would be the best, and most reliable, model for future prediction.

5.3.7 Prediction models of MDD using phenotypes and MDD PRS

Given the significant replication of the MDD prediction models using only phenotype predictors, we next build predictive models with phenotype data, but also including the MDD PRS. We follow the same procedure as described earlier, testing both main effect and interaction models. We first build predictive models using the 12 neuroticism endo-phenotypes, obesity and college phenotypes, and the MDD PRS, since the next step to improve prediction of MDD over the MDD PRS could conceivably be to add in correlated phenotype predictors. We also investigate building a predictive model using the 12 neuroticism endo-phenotypes and multiple genetic predictors, specifically the MDD, SCZ, obesity and college PRS, as we are interested in investigating whether the interaction between 'environmental' predictors, such as the neuroticism endo-phenotypes, and genetic factors can improve the prediction of MDD case-control status.

5.3.7.1 Main effect models

We fit a logistic regression model of the 12 neuroticism endo-phenotypes, obesity and college phenotypes, and MDD PRS predicting MDD case/control status, controlling for age and sex and the top 15 PCs:

$$MDD_{status} \sim MS + M + I + S + FU + N + W + T + WTL + SFN + L + G + Obese + College + MDD_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

A forward and backward variable selection procedure was applied to this full model using both AIC and BIC criterion. The best-fitting model selected based on AIC is:

$$MDD_{status} \sim MS + M + N + W + T + SFN + L + G + Obese + College + MDD_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

We compare the fit of this model to that of the null model containing only the age, sex and top 15 PCs in a likelihood ratio test (LRT), as well as to the MDD null model containing the control variables and MDD PRS. **Table 55** details the results of the LRT as well as Nagelkerke's pseudo-R^2 for this model.

Model	R^2	*P*-value
Main effects (AIC) vs. Null model	0.0617	8.12×10^{-44}
Main effects (AIC) vs. MDD Null model	0.0607	1.01×10^{-43}

Table 55. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype and MDD PRS model compared to the null model and MDD null model.

The best-fitting model selected based on BIC is:

$$MDD_{status} \sim MS + SFN + G + Age + Sex + PC_1 + \cdots + PC_{15}$$

This model contains the same phenotype predictors as the best-fitting model based on the stepwise (BIC) procedure performed earlier on the 12 neuroticism endo-phenotypes. We compare the fit of this model to that of the null model containing only the age, sex and top 15 PCs in a likelihood ratio test (LRT). **Table 56** details the results of the LRT as well as Nagelkerke's pseudo-R^2 for this model.

Model	**R^2**	***P*-value**
Main effects (BIC) vs. Null model	0.0519	3.57×10^{-42}

Table 56. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype and MDD PRS model compared to the null model.

5.3.7.2 Two-way interactions

We next fit a logistic regression model with the same predictors as in the previous main effects model, but we also include two-way interactions between the phenotypes and PRS predictors, as well as with the age and sex covariates. There are a total of 136 two-way interaction terms in addition to the 15 main effect terms and the control variables in the model, on which the variable selection procedures are performed.

The stepwise variable selection two-way interaction model as determined by AIC is:

$$
\begin{aligned}
MDD_{status} \sim\ & MS + M + I + S + FU + N + W + T + WTL + SFN + L + G \\
& + Obese + College + MDD_{PRS} + MS{:}Age + S{:}Age + L{:}Age \\
& + Age{:}College + M{:}Sex + T{:}Sex + G{:}Sex + Sex{:}Obese + MS{:}N \\
& + MS{:}G + MS{:}MDD_{PRS} + M{:}N + M{:}W + M{:}T + M{:}L + I{:}SFN \\
& + I{:}L + FU{:}SFN + FU{:}G + FU{:}College + N{:}W + T{:}SFN \\
& + WTL{:}College + WTL{:}MDD_{PRS} + G{:}College + G{:}MDD_{PRS} + Age \\
& + Sex + PC_1 + \cdots + PC_{15}
\end{aligned}
$$

We compare the fit of this model to that of the null model, the MDD null model and the stepwise (AIC) model of main effects. **Table 57** details the results of the LRTs as well as Nagelkerke's pseudo-R^2 for this model.

Model	**R^2**	***P*-value**
Interaction model (AIC) vs. Null model	0.0949	6.14×10^{-53}
Interaction model (AIC) vs. MDD null model	0.0939	1.10×10^{-52}
Interaction model (AIC) vs. Main effects (AIC) model	0.0332	4.06×10^{-14}

Table 57. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype and MDD PRS two-way interaction model compared to the null model, MDD null model and the stepwise (AIC) main effect model.

The stepwise variable selection two-way interaction model as determined by BIC is:

$$MDD_{status} \sim MS + M + N + SFN + G + MDD_{PRS} + M{:}N + G{:}MDD_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

We compare the fit of this model to that of the null model, the MDD null model and the stepwise (BIC) model of main effects. **Table 58** details the results of the LRTs as well as Nagelkerke's pseudo-R^2 for this model.

Model	**R^2**	***P*-value**
Interaction model (BIC) vs. Null model	0.0621	2.60×10^{-46}
Interaction model (BIC) vs. MDD null model	0.0611	2.75×10^{-46}
Interaction model (BIC) vs. Main effects (BIC) model	0.0103	2.09×10^{-7}

Table 58. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype and MDD PRS two-way interaction model compared to the null model, MDD null model and the stepwise (BIC) main effect model.

5.3.8 Validation of phenotype and MDD PRS models

As with the PRS predictive models and phenotype predictive models, we now aim to validate the models built in the test dataset using the phenotype data and MDD PRS.

5.3.8.1 Main effect models

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (AIC) main effect model with the null model as fitted in the validation dataset are given in **Table 59**.

Model	R^2	*P*-value
Main effects (AIC) vs. Null model	0.0645	3.92×10^{-46}
Main effects (AIC) vs. MDD null model	0.0645	7.74×10^{-47}

Table 59. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype and MDD PRS model compared to the null model and MDD null model in the validaton dataset.

Similarly, the results of the likelihood ratio test (LRT) of the BIC selected main effect model from the test dataset as fitted in the validation dataset are given in **Table 60**.

Model	R^2	*P*-value
Main effects (BIC) vs. Null model	0.0449	1.83×10^{-36}

Table 60. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype and MDD PRS model compared to the null model in the validation dataset.

We observe that both the AIC and BIC selected main effect models are significantly replicated in the independent validation dataset. These results would suggest that the AIC penalty appears to be most appropriate as the model selected based on AIC is more predictive of MDD case/control status than the model selected based on BIC.

5.3.8.2 Two-way interactions

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (AIC) interaction model with the null model, MDD null model and the stepwise (AIC) main effect model as fitted in the validation dataset are given in **Table 61**.

Model	R^2	*P*-value
Interaction model (AIC) vs. Null model	0.0735	3.31×10^{-37}
Interaction model (AIC) vs. MDD null model	0.0735	1.24×10^{-37}
Interaction model (AIC) vs. Main effects (AIC) model	0.00901	0.252

Table 61. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype and MDD PRS two-way interaction model compared to the null model, MDD null model and the stepwise (AIC) main effect model in the validation dataset.

The Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-values from the comparison of the stepwise (BIC) interaction model with the null model, MDD null model and the stepwise (BIC) main effect model as fitted in the validation dataset are given in **Table 62**.

Model	R^2	*P*-value
Interaction model (BIC) vs. Null model	0.0564	9.00×10^{-42}
Interaction model (BIC) vs. MDD null model	0.0564	1.57×10^{-42}
Interaction model (BIC) vs. Main effects (BIC) model	0.0116	2.06×10^{-8}

Table 62. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype and MDD PRS two-way interaction model compared to the null model, MDD null model and the stepwise (BIC) main effect model in the validation dataset.

We observe that the AIC selected interaction model is significant when compared to the null model and MDD null model in the validation dataset, but is not significant when compared to the stepwise AIC selected main effect model. The BIC selected interaction model significantly replicates in the validation dataset when compared to

the null model, the MDD null model and the stepwise BIC selected main effect model. These results further suggest that the BIC penalty appears to be most appropriate when considering interaction terms as the BIC selected interaction model provides additional variance explained on the BIC selected main effect model, and replicates in independent data. The AIC selected main effect model does, however, explain a greater amount of variance than the BIC selected interaction model, suggesting that the AIC main effect model would provide the greatest prediction of MDD case/control status.

5.3.9 Prediction models of MDD using phenotypes and multiple PRS

In the previous stepwise procedures, we see that building predictive models of MDD using only PRS predictors does not lead to replication in independent data, despite the individual PRS replicating. In contrast, when we use phenotype predictors only, the predictive models replicate in independent data, and when we use phenotype predictors and include the MDD PRS as a predictor, we also observe replication (though not for the AIC stepwise interaction model).

We next extend the predictive models of the previous section to use the neuroticism endo-phenotype predictors and PRS predictors for MDD, SCZ, obesity and college. We follow the same procedure as described earlier, testing both main effect and interaction models.

5.3.9.1 Main effect models

We fit a logistic regression model with the 12 neuroticism endo-phenotypes and PRS for MDD, SCZ, obesity and college predicting MDD case/control status, controlling for age and sex and the top 15 PCs:

$$MDD_{status} \sim MS + M + I + S + FU + N + W + T + WTL + SFN + L + G + MDD_{PRS} + SCZ_{PRS} + Obese_{PRS} + College_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

A forward and backward variable selection procedure is applied to this full model using both AIC and BIC criterion. The best-fitting model selected based on AIC is:

$$MDD_{status} \sim MS + M + N + W + T + SFN + L + G + MDD_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

We compare the fit of this model to that of the null model and the MDD null model. **Table 63** details the results of the likelihood ratio test (LRT) as well as Nagelkerke's pseudo-R^2 for this model.

Model	**R^2**	***P*-value**
Main effects (AIC) vs. Null model	0.0582	2.25×10^{-42}
Main effects (AIC) vs. MDD null model	0.0572	2.60×10^{-42}

Table 63. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS model compared to the null model and MDD null model.

The best-fitting model selected based on BIC is:

$$MDD_{status} \sim MS + SFN + G + Age + Sex + PC_1 + \cdots + PC_{15}$$

This model is the same BIC selected main effect model as in the previous section, as well as containing the same predictors from the BIC variable selection procedure on the phenotype predictor main effect model. We compare the fit of this model to that

of the null model and **Table 64** details the results of the LRT as well as Nagelkerke's pseudo-R^2 for this model.

Model	R^2	*P*-value
Main effects (BIC) vs. Null model	0.0519	3.57×10^{-42}

Table 64. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS model compared to the null model and MDD null model.

5.3.9.2 Two-way interactions

We next fit a logistic regression model with all the same predictors as in the previous main effects model, but we also include two-way interactions between the phenotypes and PRS predictors, and with the age and sex covariates. There are a total of 153 two-way interaction terms in the model in addition to the 16 main effect terms and the control variables, on which the variable selection procedure is performed.

The stepwise variable selection two-way interaction model as determined by AIC is:

$$
\begin{aligned}
MDD_{status} \sim\ & MS + M + I + S + FU + N + W + T + WTL + SFN + L + G \\
& + MDD_{PRS} + SCZ_{PRS} + OBESE_{PRS} + COLLEGE_{PRS} + MS{:}Age \\
& + S{:}Age + L{:}Age + Age{:}COLLEGE_{PRS} + M{:}Sex + MS{:}N \\
& + MS{:}WTL + MS{:}G + MS{:}MDD_{PRS} + M{:}N + M{:}W + M{:}T \\
& + M{:}L + M{:}SCZ_{PRS} + I{:}SFN + I{:}L + I{:}SCZ_{PRS} + I{:}COLLEGE_{PRS} \\
& + S{:}SCZ_{PRS} + S{:}OBESE_{PRS} + FU{:}SFN + FU{:}G + FU{:}SCZ_{PRS} \\
& + N{:}W + N{:}OBESE_{PRS} + W{:}OBESE_{PRS} + T{:}SFN \\
& + T{:}COLLEGE_{PRS} + WTL{:}L + WTL{:}MDD_{PRS} + L{:}SCZ_{PRS} \\
& + G{:}MDD_{PRS} + G{:}OBESE_{PRS} + MDD_{PRS}{:}OBESE_{PRS} + Age + Sex \\
& + PC_1 + \cdots + PC_{15}
\end{aligned}
$$

We compare the fit of this model to that of the null model, the MDD null model, the MDD and SCZ null model, and the stepwise (AIC) selected main effect model. The results of the LRT as well as Nagelkerke's pseudo-R^2 for this model are given in **Table 65**.

Model	R^2	*P*-value
Interaction model (AIC) vs. Null model	0.0957	2.02×10^{-49}
Interaction model (AIC) vs. MDD null model	0.0947	3.82×10^{-49}
Interaction model (AIC) vs. MDD and SCZ null model	0.0939	4.99×10^{-49}
Interaction model (AIC) vs. Main effects (AIC) model	0.0376	1.55×10^{-13}

Table 65. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS two-way interaction model compared to the null model, the MDD null model, and the stepwise (AIC) main effect model.

The best-fitting model selected based on BIC is:

$$MDD_{status} \sim MS + M + N + T + SFN + G + MDD_{PRS} + COLLEGE_{PRS} + M{:}N + T{:}COLLEGE_{PRS} + G{:}MDD_{PRS} + Age + Sex + PC_1 + \cdots + PC_{15}$$

We compare the fit of this model to that of the null model, the MDD null model and the stepwise (BIC) selected main effect model. The results of the LRT and Nagelkerke's pseudo-R^2 for this model are given in **Table 66**.

Model	R^2	*P*-value
Interaction model (BIC) vs. Null model	0.0657	5.70×10^{-47}
Interaction model (BIC) vs. MDD null model	0.0646	6.92×10^{-47}
Interaction model (BIC) vs. Main effects (BIC) model	0.0138	1.17×10^{-8}

Table 66. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS two-way interaction model compared to the null model, the MDD null model, and the stepwise (BIC) main effect model.

5.3.10 Validation of phenotype and multiple PRS models

We now rebuild the models fitted in the test dataset in the validation dataset to investigate whether they replicate in independent data.

5.3.10.1 Main effect model

We compare the fit of the AIC selected main effect model to the null model and the MDD null model as fitted in the validation dataset. The results of the LRTs and Nagelkerke's pseudo-R^2 for the model are given in **Table 67**.

Model	R^2	*P*-value
Main effects (AIC) vs. Null model	0.0617	2.84×10^{-45}
Main effects (AIC) vs. MDD null model	0.0617	5.07×10^{-46}

Table 67. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS model compared to the null model and MDD null model in the validation dataset.

We compare the fit of the BIC selected main effect model to the null model as fitted in the validation dataset. The results of the LRT and Nagelkerke's pseudo-R^2 for the model are given in **Table 68**.

Model	R^2	*P*-value
Main effects (BIC) vs. Null model	0.0449	1.83×10^{-36}

Table 68. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS model compared to the null model and MDD null model in the validation dataset.

5.3.10.2 Two-way interactions

We compare the fit of the AIC selected interaction model to the null model, the MDD null model, the MDD and SCZ null model and the stepwise AIC main effect model as

fitted in the validation dataset. The results of the LRTs and Nagelkerke's pseudo-R^2 for the model are given in **Table 69**.

Model	R^2	*P*-value
Interaction model (AIC) vs. Null model	0.0756	5.03×10^{-35}
Interaction model (AIC) vs. MDD null model	0.0756	2.06×10^{-35}
Interaction model (AIC) vs. MDD and SCZ null model	0.0748	2.84×10^{-35}
Interaction model (AIC) vs. Main effects (AIC) model	0.0139	0.0909

Table 69. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (AIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS two-way interaction model compared to the null model, the MDD null model, and the stepwise (AIC) main effect model in the validation dataset.

We compare the fit of the BIC selected interaction model to the null model, the MDD null model, and the stepwise BIC main effect model as fitted in the validation dataset. The results of the LRTs and Nagelkerke's pseudo-R^2 for the model are given in **Table 70**.

Model	R^2	*P*-value
Interaction model (BIC) vs. Null model	0.0594	5.09×10^{-42}
Interaction model (BIC) vs. MDD null model	0.0594	1.05×10^{-42}
Interaction model (BIC) vs. Main effects (BIC) model	0.0145	3.44×10^{-9}

Table 70. Nagelkerke's pseudo-R^2 and likelihood ratio test (LRT) *P*-value for the stepwise (BIC) neuroticism endo-phenotype, MDD, SCZ, Obese and College PRS two-way interaction model compared to the null model, the MDD null model, and the stepwise (BIC) main effect model in the validation dataset.

Here we see the same pattern of replication as in the previous section. The AIC and BIC selected main effect models both replicate in the independent validation dataset. However, for the interaction models, the AIC and BIC models are significant compared to the null models, but only the BIC model provides significantly better prediction over the corresponding main effect model in the validation dataset. That said, the AIC main effect model is again the model that provides the greatest

variance explained out of all the replicated models, suggesting that the AIC selected main effect model would offer the greatest, and most reliable, prediction of MDD case/control status.

5.4 Discussion

In this study we applied stepwise variable selection procedures to logistic regression with multiple predictors, both genetic (PRS) and phenotypic, in order to establish a prediction model of MDD case/control status from UK Biobank data. We used both AIC and BIC criterion for variable selection in order to compare the approaches, and considered two-way interaction terms between our predictors.

We built PRS for MDD, SCZ and for MDD comorbid traits: neuroticism (total score and the 12 endo-phenotypes), obesity and whether college was attended. Previous analyses into the predictive power of the MDD PRS have yielded disappointing results, and the SCZ PRS has even been shown to be a better predictor of MDD case/control status (Euesden et al., 2015). We therefore explored the use of multiple PRS as predictors of MDD case/control status. In order to investigate whether the interaction between the genetics of, for example, MDD and obesity associates with MDD case/control status, we considered the interaction between PRS. While the results from the prediction models in the test dataset seemed promising with greater phenotypic variance explained than the MDD PRS alone, the models did not replicate in the independent validation dataset. This suggests that both the AIC and BIC stepwise variable selection procedures produced models that were over-fit to the test data. Particularly with the two-way interaction terms, there is a substantial possibility that there are low numbers of individuals to which each interaction is applicable, thus potentially leading to over-fitting of the model. Furthermore, given such large sample sizes in the UK Biobank even small effects potentially have the power to be detected, which then have a low probability of being validated in independent data, since validation datasets are likely to be of a much smaller sample size and thus potentially underpowered to detect the same magnitude of effect.

Given that there are known phenotypic correlations between MDD and neuroticism, obesity and educational attainment (Smith et al., 2016; Okbay et al., 2016; Hung et al., 2015; Rivera et al., 2012; Bjelland et al., 2008), we also performed the same variable selection procedures using only the phenotype data. In contrast to the PRS predictive models, the phenotype main effect models (both AIC and BIC selected) did replicate in an independent dataset. While the AIC selected two-way interaction model did not replicate, the BIC selected two-way interaction model did. This suggests that using the more stringent BIC to determine the predictors to be included in the model, particularly when considering interaction terms, may yield a more robust predictive model with greater probability of replication in an independent dataset. Though generally the models resulting from BIC variable selection had lower pseudo-R^2 values than the AIC selected equivalents, meaning that they explained less variance in MDD case/control status, the BIC selected models are likely to be more widely applicable.

Finally, we investigated the use of a combination of phenotype and genetic predictors of MDD case/control status. For both the models using phenotypes and MDD PRS predictors, and the models using phenotypes, MDD, SCZ, obesity and college PRS predictors, we observed replication for both the AIC and BIC main effect models, while only the BIC interaction models replicated.

From our investigation of building predictive models of MDD case/control status using PRS predictors and phenotype predictors, there was much greater replication of the predictive models when using the phenotype data. Given that the MDD PRS itself provides only modest prediction of MDD case/control status, building a predictive model of MDD using PRS of correlated traits is likely to be less powerful than using only the phenotypic data due to the genetic heterogeneity of MDD. Although our attempts to build a predictive model of MDD by considering $PRS \times PRS$

interactions did not replicate, we did achieve replication for models including $PRS \times environment$ and $environment \times environment$ interaction terms. In the two-way interaction (BIC) phenotype predictor model, we observed interactions of the fed-up feelings and worry too long phenotypes with the college phenotype; in the two-way interaction (BIC) phenotype and MDD PRS model, we retained interaction terms between miserableness and nervous feelings, and guilty feelings and the MDD PRS; finally, in the two-way interaction (BIC) phenotype and multiple PRS predictor model we observed interactions between the miserableness and nervous feelings phenotypes, the tense feelings phenotype and the college PRS, and the guilty feelings phenotype and the MDD PRS. Within genetic epidemiology, most research into interactions has involved the testing and discovery of $genetic\ variant \times environment$ interactions (Caspi et al., 2003), but there has been minimal focus on investigating how the genome-wide genetic burden interacts with the environment (Mullins et al., 2016; Keers et al., 2016). The findings presented here provide an indication that these interactions may in fact often exist and contribute to the variance in human phenotypes. Therefore, this work highlights the need for further investigation into this, starting with replication using different phenotypes and data, to explore what impact interactions between genome-wide genetic burdens for certain traits and environmental risk factors have on physical disease and psychiatric disorders.

Considering future developments in population-level genotyping and the growing surge towards 'precision medicine', it is likely that predictive models of disease will become more frequently developed and applied, which provided the motivation for this study. Genetic risk score approaches have been shown to provide improved prediction over existing predictors of type 2 diabetes risk (Läll et al., 2016), demonstrating the utility of this approach for the prediction of disease risk. The results from the phenotype data prediction models suggest that there is potentially

statistical power to be gained from performing multi-trait analyses on components and symptoms of major phenotypes rather than using binary summaries of them. However, the general challenge in predicting MDD from genetic data may have resulted in insufficient power in this study to demonstrate directly the benefit of multi-trait PRS prediction of MDD. Although our PRS-only predictive models did not replicate in an independent dataset, if the same procedure were applied to a more genetically homogeneous phenotype we may obtain more promising results.

6. Discussion

We have presented a variety of research developments and applications under the multiple phenotype theme, ranging from adjustments of univariate GWAS summary statistics to obtain multi-trait association *P*-values, to building prediction models of disease that incorporate genetic and environmental interactions. In each instance, we have demonstrated the utility in taking a multivariate approach to statistical modelling, taking care to account for the complexities introduced when transitioning from single-phenotype analyses. Overall, our findings suggest that taking a multivariate approach yields increased statistical power and greater understanding of the relationships underlying multiple correlated traits than single phenotype analyses alone. Considering phenotypes jointly can expose previously unknown connections that can aid in shaping the direction of future analyses, and bring us closer to the ultimate goal of understanding how the complex biological network underlying multiple diseases and traits operates.

6.1 Multivariate simulation framework

In **Chapter 2**, we set up a modelling framework for exploring the implications of modelling multiple phenotypes in the context of genome-wide association studies (GWAS). Motivated by the recent developments in multi-trait GWAS methodology and the lack of understanding into their relative performance in the field, we developed a simulation framework for generating multi-trait GWAS data with known causal genetic relationships for the benchmarking of multi-trait GWAS methods. When modelling only one trait with a causal genetic relationship relating to one SNP, the important considerations for genetic association testing are the sample size, the minor allele frequency of the SNP and the magnitude (and direction) of the genetic

effect from the SNP to the trait. Modelling multiple traits, however, produces extra dimensionality: the phenotypic correlation structure, and the interplay between these correlations and the combination of genetic effects on the traits.

We developed a series of simulation scenarios that aimed to capture a large portion of the multivariate data landscape, in order to expose the similarities and differences between current multi-trait GWAS methodology. We built simulation scenarios where the genetic effects and phenotypic correlations were varied in a structured way, where they were varied more freely, and where the phenotypic correlations were reflective of the genetic effects on the traits. To perform simulations that closely matched reality, we also exploited publicly available summary data from univariate GWAS to inform the genetic effects, and used phenotypic correlations extracted directly from the Northern Finland Birth Cohort (NFBC1966). The main data-generating model considered direct effects between the SNP and multiple quantitative traits, but we also modelled indirect genetic effects as well as binary traits.

The simulation framework developed here is available as an open-source command line program to act as an aid in method development and benchmarking. We also provide an R shiny web-application (www.MultiTraitGWAS.kcl.ac.uk) that generates multivariate datasets that can be used for applications similar to those presented here. With minor modifications, data can be generated using our open-source software for any application concerning multiple correlated variables with some common associated factor.

While we aimed to make the simulation scenarios implemented in our framework as extensive as possible, it inevitably could not be exhaustive, especially for multivariate simulations where there are infinite combinations of phenotypic correlations and

genetic effects. These limitations formed part of the motivation for creating the associated software package, so that it can be utilised by researchers in the future, and scenarios can be expanded upon as required. We also did not consider the simulation of correlated SNP data in linkage disequilibrium (LD). The focus of this study was single-SNP, multi-trait methods in order to isolate the methodological advances made by modelling multiple traits jointly, in comparison to the standard univariate approach. However, the software could be modified to model such data, and the performance of multi-SNP, multi-trait methods can thus be compared in a similar way.

6.2 Multi-trait GWAS methods comparison

The focus of **Chapter 3** was to perform a comparison of the leading multi-trait GWAS methods by utilising the simulation framework developed in **Chapter 2**. Previous publications introducing new methodology (O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015; Aschard et al., 2014), or previous comparison studies (Galesloot et al., 2014), considered only a few methods or a small number of simulation scenarios, making it difficult to fully dissect and contrast the performance of existing methodology. This provided the motivation to develop a rigorous set of simulation scenarios, presented in **Chapter 2**, to fully test the performance of the methodology, thereby demystifying user choice and enabling higher-powered multi-trait analyses in the future.

In this comparison study, we compared the leading multi-trait GWAS methodology, consisting of two types of method: those that exploit existing GWAS summary statistics (O'Reilly et al., 2012; van der Sluis et al., 2013; Zhu et al., 2015) and those that utilise individual-level genotype-phenotype data (Ferreira and Purcell, 2009;

Nath and Pavur, 1985; O'Reilly et al., 2012; Aschard et al., 2014; Stephens, 2013; Marchini et al., 2007). Ten methods were compared in total, though further methods can be easily incorporated into our open-source software package. These methods were compared across all four scenarios implemented in the simulation framework, for varying numbers of traits, genetic effects and phenotypic correlations.

The results of the comparison study suggest that the methods that utilise individual-level genotype-phenotype data will, generally, optimise discovery power, based on the final, most realistic, scenario in the simulation framework where it was estimated that individual-level methods could yield twice the discovery of genetic variants over the summary statistic methods. However, summary data are likely to be available on much larger sample sizes than available in multi-trait GWAS panels, in particular in relation to case/control data. From simulations at a larger sample size for the summary statistic methods, we observed a substantial increase in power over the individual-level methods, suggesting that if summary statistics are available on much larger sample sizes then these methods will provide optimal discovery power. That said, the emergence of population-wide resources such as the UK Biobank means that multi-trait analyses on individual-level data will soon be able to match, and even exceed, the sample sizes of summary data, making the development of individual-level multi-trait methods still an important area of research. As sample sizes grow, the main consideration in methodology development will be in reducing computation time to ensure methods are computationally feasible for application in these large population cohorts, especially since our results suggest that optimal power for multi-trait methods has already been reached.

6.3 Summary statistic GWAS

Motivated by the results from our comparison study, where we observed that summary statistic GWAS methods could often achieve similar discovery power as methods that exploit individual-level data, and greater power with larger data publicly available, we performed a series of multi-trait analyses on publicly available GWAS summary statistics. From simulations of scenario S4b on two traits, as presented in **Chapter 3**, we observe in **Figure 45** that with only modest gains in sample size the summary statistic method of S_{Het} can achieve power gains over the individual-level approach.

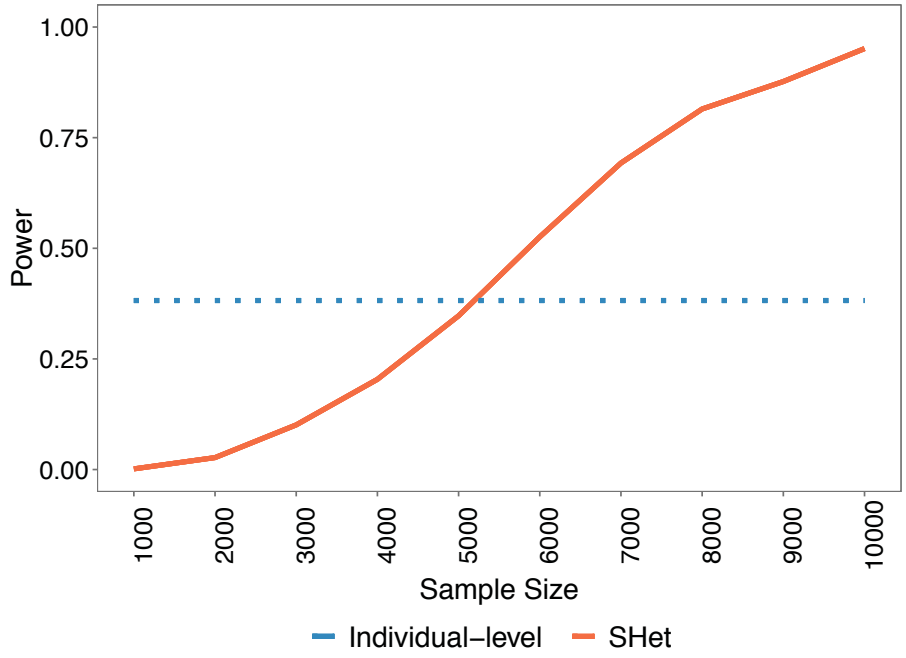

Figure 45. Power of the summary statistic method S_{Het} from simulations of scenario S4b on two traits for varying sample sizes – the dotted line represents the power of the individual-level multi-trait GWAS methods for $5{,}000$ samples.

In recent years the availability of GWAS summary data has increased, though up to this point the main utility for these data has been in meta-analyses of the same trait in order to boost discovery power (Ripke et al., 2013; Hyde et al., 2016; Morris et al.,

2012; Global Lipids Genetics Consortium, 2013), as well as for assessing the shared genetic aetiology between traits using polygenic risk scores (PRS) (Purcell et al., 2009; Euesden et al., 2015; Dudbridge, 2013; Selzam et al., 2016; Krapohl et al., 2015; Power et al., 2015) and the LD Score regression method (B. Bulik-Sullivan et al., 2015; B. K. Bulik-Sullivan et al., 2015). Multi-trait methods that exploit summary data to perform meta-analyses across traits have now been developed (Zhu et al., 2015; Bolormaa et al., 2014; Zhu and Stephens, 2016), enabling existing data to be repurposed to boost the power to detect causal variants through the joint analysis of multiple correlated traits.

We collected summary statistics from the largest GWAS to date on 19 traits, consisting of anthropometric, metabolic and psychiatric phenotypes. Next we developed our own summary statistic based multi-trait GWAS methods, MetaHom and MetaHet, introduced in this thesis, by extending existing methods to make them applicable to a mixture of quantitative and binary traits, while allowing for opposing directions of genetic effect on different traits (Zhu et al., 2015). Here we focused on the application of these two methods in particular, because the results of a comparison of several summary statistic methods that we conducted showed that MetaHom has greatest power when pleiotropic genetic effects exist, and MetaHet has highest power when heterogeneous genetic effects exist. By using a combination of these two methods, we expect to maximise our potential to uncover novel causal variants when the underlying genetic aetiology of the traits is unknown prior to the analyses. Multi-trait GWAS were performed on 16 correlated sets of traits, as well as on all 19 traits jointly. Novel associations were identified across the analyses, which we validated as being highly likely to be genuine signals via evaluation of their CADD scores.

Since performing this study, summary data on many more phenotypes has been made publicly available, meaning that the potential for further multi-trait analyses has expanded. The *PhenoScanner* online tool (Staley et al., 2016) will, in addition, facilitate the pooling of summary data across many traits, further aiding the utilisation of such data. Furthermore, with the release of the UK Biobank data, there is likely to be vast amounts of summary data resulting from univariate studies of this resource, aiding follow-up analyses and potentially leading to novel discoveries as demonstrated here. While we performed analyses on many traits, the phenotypes under study in this chapter are likely to be quite homogenous, for example the lipids and obesity measures. Thus, further multi-trait analyses across many more phenotypes will help to further paint the picture as to the performance of multi-trait summary statistic GWAS methods.

This study highlights the utility of summary statistic data, and demonstrates that additional signal can be extracted by multi-trait analyses, and by applying the most appropriate summary statistic method for identifying genotype-phenotype relationships. We can also gain further insight into the biology underlying multiple traits from such analyses; for example, MetaHom is highly powered to detect pleiotropic genetic effects, and so any genetic variant identified by the application of this method is likely to be a pleiotropic SNP, thus providing insight beyond that obtained from univariate GWAS.

6.4 Prediction modelling using PRS

An ultimate aim of genetics research is to understand the human genome sufficiently to be able to predict phenotypes given an individual's DNA. If we are able to completely understand the biological network, its interactions with internal and

external factors, and the link between genetic variation and phenotypic outcomes then, in theory, we could build a highly powerful predictive model for any disease outcome. The field, however, is not at this point yet and as it stands there is still a large amount of phenotypic variance and heritability left unexplained. Polygenic risk scores (PRS), however, have shown signs of success for the prediction of diseases and traits from information on genetic risk of disease (Selzam et al., 2016; Krapohl et al., 2015; Power et al., 2015; Vassos et al., 2016). The most successful prediction of a behavioural trait using PRS has been in educational attainment, where the PRS for educational attainment was found to explain 9% of the variance in educational achievement at age 16 (Selzam et al., 2016). The twin heritability of educational achievement has been estimated to be around 60% (Krapohl et al., 2014), and so the upper limit on the amount of variance a highly-powered PRS could explain is around this figure. While 9% may seem rather modest, this compares to $\sim 1\%$ of the variance explained by sex, and shows that the field is making real progress as larger sample sizes are obtained and as studies become more powerful. In contrast, similar success has not been observed for phenotypes such as MDD; the MDD PRS currently only explains 0.6% of the variance in MDD case/control status (Ripke et al., 2013). Most PRS prediction models have been performed either within phenotype or between two phenotypes to assess their shared genetic aetiology (Purcell et al., 2009; Euesden et al., 2015; Selzam et al., 2016). Given the multi-dimensionality of heterogeneous traits, taking a multivariate approach to prediction could yield greater success.

In **Chapter 5** we built predictive models of MDD using multiple PRS predictors computed from the UK Biobank, as well as comparing this approach to building predictive models using only phenotype data. We considered both the main effects and two-way interactions between the predictors in order to establish the most

predictive model of MDD. We built models consisting of only PRS predictors, only phenotype predictors, and of a combination of PRS and phenotype predictors, implementing variable selection procedures to determine the most predictive models (AIC and BIC). We then replicated the models in an independent subset of the UK Biobank. We found that, although the PRS built in the validation dataset significantly predicted the trait that they were built on, the prediction models of MDD did not replicate, for both the main effects and the interaction models. Replication, however, was observed for the models using only phenotype predictors and the models using a combination of phenotypes and PRS, both for the main effects models and the BIC selected interaction models. There was no instance where the AIC selected interaction model replicated, suggesting that these models were over-fit to the test dataset. This also suggests that when considering interaction terms on these type of data, BIC may be the most appropriate variable selection criterion to prevent over-fitting and improve the generalisability of the predictive model. When considering the interaction between different PRS, interaction terms could be retained in a model yet apply to only a small number of individuals in the dataset, thus leading to over-fitting. Furthermore, the size of the UK Biobank resource means there is sufficient power to detect small effects that are very specific to a subgroup of the total population. These particular issues are amplified when studying such a heterogeneous disorder as MDD, as there is likely to exist many subtypes that can be highlighted by interaction terms, but not be widely applicable to MDD as a whole. This approach applied to a more homogenous trait may achieve greater success.

Even though the multiple PRS-only models did not replicate, we did achieve success using the phenotype data alone, as well as the models built using a combination of PRS and phenotype predictors. One of the phenotypes of particular interest in this study was neuroticism, due to its known association with MDD and their high phenotypic correlation (Smith et al., 2016; Kendler and Myers, 2010). The

neuroticism score is made up of 12 components, assessing different aspects of neuroticism. We performed analyses to assess whether greater prediction of MDD could be achieved from using these component measures as multiple predictors, instead of the one aggregated neuroticism score. We performed both AIC and BIC variable selection on the model with all 12 neuroticism components as predictors, and both of the resulting models offered greater prediction of MDD; the neuroticism score model explained 4.8% variance in MDD, whereas the AIC selected neuroticism component model explained 5.8% variance in MDD. By decomposing traits into their components, we can not only improve prediction but also gain greater understanding into the biology of the traits. The AIC variable selection procedure retained only a subset of the neuroticism components in the most predictive model, providing insight into the components that most associate with MDD. This information can be used to further guide the analysis of MDD as a phenotype, and highlight important factors for phenotyping and subtype analyses. Multivariate analyses, such as those presented here, could hold the key to understanding the genetic aetiology of heterogeneous disorders.

6.5 Future work

The field of genetic epidemiology is evolving rapidly. With new methodology being developed and new discoveries being made at a rapid pace, both now in relation to genetic association and genetic prediction, researchers must work dynamically as we learn more and more about the human genome. Due to such advances in understanding, there is always more to uncover with follow up analyses. One question answered can lead to the creation of several more as we delve further into the analysis of the genetic determinants of disease.

6.5.1 Multi-SNP simulation and methods comparison

Having developed a simulation framework as a platform for the comparison of multi-trait, single-SNP GWAS methods, the next step is to extend this framework to incorporate the simulation of multiple SNPs associated with multiple phenotypes. The focus of multi-SNP methods (Bottolo et al., 2013; Zhou and Stephens, 2012; Kim et al., 2016) is to gain additional power by considering the association between sets of SNPs and multiple traits by reducing residual variation. Simulation of independent SNPs associated with multiple traits would be a simple extension of the main data-generating model of **Chapter 2.** Instead here, for example, we may generate, say, 100 SNPs each explaining 0.01% of the variance in each trait (thus $h^2 = 1\%$ in total), ensuring that the residual error term accounts for 99% of the phenotypic variance. This holds under the assumption of small total h^2, but to simulate SNPs of larger total effect we would not be able to continue to approximate the phenotypic correlation by the residual correlation and instead would need to construct a proper, non-approximate, model using conditional expectations via application of Bayes' formula. Alternatively, vast individual-level data resources, such as the UK Biobank,

could also be exploited to generate a large amount of highly realistic simulation data, for example by selecting SNPs at random and constructing traits by assuming a causal effect of the SNP(s) with certain heritability. The performance of multi-SNP, multi-trait methods can then be explored and help to guide the future direction of GWAS methodology. Given that there is a highly connected network of genetic interactions underlying most common diseases and traits, multi-SNP methods have the potential to uncover further insight into how these genetic interactions influence human diseases. As technology advances and modelling interactions between millions of SNPs across the genome becomes more computationally feasible, multi-SNP methods will have an important role to play in the analysis of the effect genetic variation has on phenotypic outcomes.

6.5.2 Multi-trait GWAS in the UK Biobank

The results of our comparison of multi-trait GWAS methods suggested that, taking sample size into account, summary statistic GWAS methods were likely to provide the greatest discovery power. This provided the motivation for the application of summary statistic GWAS methods to publicly available summary data in **Chapter 4**. The application of these methods led to the identification of novel associations, which the separate univariate analyses did not have power to detect. However, the emergence of the UK Biobank resource, which has currently collected genotype data on around $140{,}000$ individuals and is set to release genotype data on the full sample (around $500{,}000$ individuals) in January 2017, will facilitate highly powered multi-trait analyses of individual-level data, exceeding the sample sizes of available summary data. The results of our simulations suggest that applying an individual-level data method on these data would optimise discovery power. We plan to perform multi-trait analyses on this resource using a range of individual-level methods, to act as a

further comparison of these methods on real data, but also to boost discovery power for the identification of causal genetic variants. We are particularly interested in building upon our analyses presented here to help leverage underpowered phenotypes, such as MDD, in order to further understand the genetic aetiology of heterogeneous disorders. An important challenge in the future will be fully exploiting both multi-trait individual-level resources, such as the UK Biobank, and summary data available from large-scale GWAS meta-analyses simultaneously, rather than using only one resource; multi-trait methods will be required to perform multi-trait analyses on each data type in a way that optimises power for each, and then appropriately combine the two sets of results to produce an overall result.

6.5.3 Prediction modelling

As the understanding of the human genome increases, prediction modelling will become an important area of research in genetic epidemiology, and research has been performed into the predictive utility of polygenic models (Chatterjee et al., 2013, 2016). Relatively modest prediction has been achieved to date for heterogeneous disorders, such as MDD (Ripke et al., 2013), though large resources of genetic data will help to facilitate this. From our study focused on building a predictive model of MDD we were able to significantly predict MDD using only phenotype data, or a combination of phenotypes and genetic factors (PRS). However, the predictive models built only on PRS did not replicate in an independent validation dataset. We concluded that due to the heterogeneity of MDD, variable selection procedures on the PRS predictors were more likely to lead to over-fitting of the prediction models, though there are several other factors that could be contributing to lack of power. Diagnosis of MDD, for example, is challenging due to varying severities of the disorder, and the stigma associated with mental health and the barriers this causes in seeking medical help. In addition, while the pooling of individuals with mild single

episode depression with those who have lifelong problems will increase the sample size, this may not be the optimum study design, and it could be that treating them as two separate disorders may lead to increased power. As larger resources of data are collected, stratified analyses can be performed to identify more homogenous clinical sub-types. Given that our genetic prediction models did not replicate, we plan to apply the same approach to more homogenous disorders and those that have already shown reasonable predictive power from genetics, such as educational attainment, schizophrenia and obesity, to further explore the utility of multiple PRS predictors in predicting disease status.

6.5.4 Rare variant analyses

The GWAS study design has been shown to be a successful tool for the identification of common variants associated with complex disorders, yet despite the identification of thousands of genetic variants there still remains the problem of missing heritability (Manolio et al., 2009; Eichler et al., 2010; Lee et al., 2011). For type 2 diabetes, for example, GWAS have identified more than 70 genetic loci, yet only $\sim 11\%$ of the heritability is explained by these common variants (Morris et al., 2012; Lee et al., 2014). Rare variants play an important role in Mendelian disorders, and there is now empirical evidence that rare variants could contribute to complex disorders (Rivas et al., 2011; Gudmundsson et al., 2012). The cost of whole-genome sequencing (WGS) now makes it possible to explore the effects of rare variants, but performing genome-wide analyses of WGS data poses many statistical challenges and questions. There are many additional factors that come with WGS, one such example being read-depth, and the trade-off between greater depth and larger sample size (Pasaniuc et al., 2012). In terms of methodology, the challenges arise in applying a method that can detect variants of low frequency. In theory the single variant design, as is

implemented in current GWAS, could be applicable for the study of rare variants. However, large sample sizes would be required in order to have sufficient statistical power, which in turn incurs extra cost. Instead a region or gene based approach is likely to yield greater success, based on the assumption that a locus containing one causal variant is likely to contain other causal variants nearby, as detailed in a review of current rare variant methodology (Lee et al., 2014). In the same way, a genetic variant or locus affecting one phenotype is likely to affect another correlated phenotype, as evidenced by recent research into genome-wide pleiotropy (Visscher and Yang, 2016; Solovieff et al., 2013). Therefore, multi-trait approaches could be even more beneficial in the analysis of rare variants, for which this sharing of information across different variants and traits may make the critical difference in power to produce discovery. As WGS becomes more frequently performed, the associated methodology should be rigorously benchmarked, as presented here, for multi-trait GWAS methods, in order to maximise the discovery potential of these data.

References

A. P. Dawid (1981) *Some matrix-variate distribution theory: notational considerations and a bayesian application*. Vol. Biometrika, 68:265-274.

Adhikari, K. et al. (2016) A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nature Communications*. [Online] 711616.

Adhikari, K. et al. (2015) A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nature Communications*. [Online] 67500.

Anon (Fact Sheet) *WHO | Depression* [online]. Available from: http://www.who.int/mediacentre/factsheets/fs369/en/ (Accessed 24 September 2016).

Aschard, H. et al. (2014) Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies. *American Journal of Human Genetics*. [Online] 94 (5), 662–676.

Berndt, S. I. et al. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics*. [Online] 45 (5), 501–512.

Bjelland, I. et al. (2008) Does a higher educational level protect against anxiety and depression? The HUNT study. *Social Science & Medicine*. [Online] 66 (6), 1334–1345.

Bland, J. M. & Altman, D. G. (1995) Multiple significance tests: the Bonferroni method. *BMJ : British Medical Journal*. 310 (6973), 170.

Bolormaa, S. et al. (2014) A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. *PLoS Genet*. [Online] 10 (3), e1004198.

Bottolo, L. et al. (2013) GUESS-ing Polygenic Associations with Multiple Phenotypes Using a GPU-Based Evolutionary Stochastic Search Algorithm. *PLoS Genet*. [Online] 9 (8), e1003657.

Bottolo, L. & Richardson, S. (2010) Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*. [Online] 5 (3), 583–618.

Bulik-Sullivan, B. et al. (2015) An atlas of genetic correlations across human diseases and traits. *Nature Genetics*. [Online] advance online publication. [online]. Available from: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3406.html (Accessed 26 October 2015).

Bulik-Sullivan, B. K. et al. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*. [Online] 47 (3), 291–295.

Burton, P. R. et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. [Online] 447 (7145), 661–678.

Casale, F. P. et al. (2015) Efficient set tests for the genetic analysis of correlated traits. *Nature Methods*. [Online] 12 (8), 755–758.

Caspi, A. et al. (2003) Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science*. [Online] 301 (5631), 386–389.

Chang, C. C. et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. [Online] 47.

Chatterjee, N. et al. (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*. [Online] 17 (7), 392–406.

Chatterjee, N. et al. (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*. [Online] 45 (4), 400–405.

Converge Consortium (2015) Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*. [Online] 523 (7562), 588–591.

Cornelis, M. C. et al. (2015) Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. Molecular psychiatry. [Online] 20 (5), 647–656.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*. [Online] 45 (9), 984–994.

Cross-Disorder Group of the Psychiatric Genomics Consortium & Genetic Risk Outcome of Psychosis (GROUP) Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*. [Online] 381 (9875), 1371–1379.

Davey Smith, G. & Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics. [Online] 23 (R1), R89–R98.

Dudbridge, F. (2013) Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genet*. [Online] 9 (3), e1003348.

Dupuis, J. et al. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*. [Online] 42 (2), 105–116.

Eichler, E. E. et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*. [Online] 11 (6), 446–450.

Euesden, J. et al. (2015) PRSice: Polygenic Risk Score software. *Bioinformatics*. [Online] 31 (9), 1466–1468.

Eysenck, S. et al. (1985) A revised version of the psychoticism scale. *Pers Indiv Differ*. 621–29.

Falconer, D. S. (1960) *Introduction to quantitative genetics*. New York,: Ronald Press Co. [online]. Available from: http://archive.org/details/introductiontoq00falc (Accessed 26 October 2015).

Feitosa, M. F. et al. (2013) The ERLIN1-CHUK-CWF19L1 gene cluster influences liver fat deposition and hepatic inflammation in the NHLBI Family Heart Study. Atherosclerosis. [Online] 228 (1), 175–180.

Ferreira, M. A. R. & Purcell, S. M. (2009) A multivariate test of association. *Bioinformatics*. [Online] 25 (1), 132–133.

Franke, A. et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics. [Online] 42 (12), 1118–1125.

Galesloot, T. E. et al. (2014) A Comparison of Multivariate Genome-Wide Association Methods. *PLOS ONE*. [Online] 9 (4), e95923.

Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2 (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nature Genetics. [Online] 42 (11), 985–990.

Gieger, C. et al. (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature*. [Online] 480 (7376), 201–208.

Global Lipids Genetics Consortium (2013) Discovery and refinement of loci associated with lipid levels. *Nature Genetics*. [Online] 45 (11), 1274–1283.

Guan, Y. & Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*. [Online] 5 (3), 1780–1815.

Gudmundsson, J. et al. (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics*. [Online] 44 (12), 1326–1329.

Han, B. et al. (2016) A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nature Genetics*. [Online] 48 (7), 803–810.

Han, B. & Eskin, E. (2011) Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *American Journal of Human Genetics*. [Online] 88 (5), 586–598.

Horikoshi, M. et al. (2013) New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature Genetics*. [Online] 45 (1), 76–82.

Huang, J. et al. (2011) PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics*. [Online] 27 (9), 1201–1206.

Hung, C.-F. et al. (2015) A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder. *BMC medicine*. [Online] 1386.

Hyde, C. L. et al. (2016) Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics*. [Online] advance online publication. [online]. Available from: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3623.html (Accessed 23 August 2016).

Hyman, S. (2014) *Mental health: Depression needs large human-genetics studies : Nature News & Comment* [online]. Available from: http://www.nature.com/news/mental-health-depression-needs-large-human-genetics-studies-1.16300 (Accessed 26 October 2015).

Jylhä, P. & Isometsä, E. (2006) The relationship of neuroticism and extraversion to symptoms of anxiety and depression in the general population. *Depression and Anxiety*. [Online] 23 (5), 281–289.

Kauwe, J. S. K. et al. (2014) Genome-Wide Association Study of CSF Levels of 59 Alzheimer's Disease Candidate Proteins: Significant Associations with Proteins Involved in Amyloid Processing and Inflammation. *PLoS Genet*. [Online] 10 (10), e1004758.

Keers, R. et al. (2016) A Genome-Wide Test of the Differential Susceptibility Hypothesis Reveals a Genetic Predictor of Differential Response to Psychological Treatments for Child Anxiety Disorders. *Psychotherapy and Psychosomatics*. [Online] 85 (3), 146–158.

Kendler, K. S. & Myers, J. (2010) The genetic and environmental relationship between major depression and the five-factor model of personality. *Psychological Medicine*. [Online] 40 (5), 801–806.

Kim, J. et al. (2016) Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. *Genetics*. [Online] 203 (2), 715–731.

Kircher, M. et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*. [Online] 46 (3), 310–315.

Klei, L. et al. (2008) Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*. [Online] 32 (1), 9–19.

Korte, A. et al. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*. [Online] 44 (9), 1066–1071.

Krapohl, E. et al. (2015) Phenome-wide analysis of genome-wide polygenic scores. *Molecular Psychiatry*. [Online] [online]. Available from: http://www.nature.com/mp/journal/vaop/ncurrent/full/mp2015126a.html (Accessed 26 October 2015).

Krapohl, E. et al. (2014) The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proceedings of the*

National Academy of Sciences of the United States of America. [Online] 111 (42), 15273–15278.

Läll, K. et al. (2016) Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine*. [Online] [online]. Available from: http://www.nature.com/gim/journal/vaop/ncurrent/full/gim2016103a.html#results (Accessed 28 September 2016).

Lambert, J.-C. et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*. [Online] 45 (12), 1452–1458.

Lee, S. et al. (2014) Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*. [Online] 95 (1), 5–23.

Lee, S. H. et al. (2011) Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics*. [Online] 88 (3), 294–305.

Lee, S. H. et al. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics (Oxford, England)*. [Online] 28 (19), 2540–2542.

Levine, M. E. et al. (2014) A polygenic risk score associated with measures of depressive symptoms among older adults. *Biodemography and Social Biology*. [Online] 60 (2), 199–211.

Li, M.-X. et al. (2011) GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *American Journal of Human Genetics*. [Online] 88 (3), 283–293.

Lippert, C. et al. (2011) FaST linear mixed models for genome-wide association studies. *Nature Methods*. [Online] 8 (10), 833–835.

Locke, A. E. et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature. [Online] 518 (7538), 197–206.

Ma, C. et al. (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology*. [Online] 37 (6), 539–550.

Manolio, T. A. et al. (2009) Finding the missing heritability of complex diseases. *Nature*. [Online] 461 (7265), 747–753.

Marchini, J. et al. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*. [Online] 39 (7), 906–913.

Milaneschi, Y. et al. (2016) Polygenic dissection of major depression clinical heterogeneity. *Molecular Psychiatry*. [Online] 21 (4), 516–522.

Morris, A. P. et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*. [Online] 44 (9), 981–990.

Mullins, N. et al. (2016) Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychological Medicine*. [Online] 46 (4), 759–770.

Nath, R. & Pavur, R. (1985) A New Statistic in the One-way Multivariate Analysis of Variance. *Comput. Stat. Data Anal.* [Online] 2 (4), 297–315.

Nieuwboer, H. A. et al. (2016) GWIS: Genome-Wide Inferred Statistics for Functions of Multiple Phenotypes. *The American Journal of Human Genetics*. [Online] 0 (0), . [online]. Available from: http://www.cell.com/ajhg/abstract/S0002-9297(16)30321-4 (Accessed 24 September 2016).

Nyholt, D. R. (2004) A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. *American Journal of Human Genetics*. 74 (4), 765–769.

Okbay, A. et al. (2016) Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*. [Online] 48 (6), 624–633.

O'Reilly, P. F. et al. (2012) MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS ONE*. [Online] 7 (5), e34861.

Palla, L. & Dudbridge, F. (2015) A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. *The American Journal of Human Genetics*. [Online] 97 (2), 250–259.

Park, J.-H. et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*. [Online] 42 (7), 570–575.

Pasaniuc, B. et al. (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*. [Online] 44 (6), 631–635.

Pasaniuc, B. & Price, A. L. (2016) Dissecting the genetics of complex traits using summary association statistics. *bioRxiv*. [Online] 72934.

Pickrell, J. K. et al. (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*. [Online] 48 (7), 709–717.

Pirinen, M. et al. (2013) Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*. [Online] 7 (1), 369–390.

Power, R. A. et al. (2015) Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience*. [Online] 18 (7), 953–955.

Price, A. L. et al. (2011) Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLOS Genet*. [Online] 7 (2), e1001317.

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*. [Online] 43 (10), 977–983.

Purcell, S. M. et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. [Online] 460 (7256), 748–752.

Reich, D. E. et al. (2001) Linkage disequilibrium in the human genome. *Nature*. [Online] 411 (6834), 199–204.

Reich, D. E. & Lander, E. S. (2001) On the allelic spectrum of human disease. *Trends in Genetics*. [Online] 17 (9), 502–510.

Rietveld, C. A. et al. (2013) GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science*. [Online] 340 (6139), 1467–1471.

Ripke, S. et al. (2013) A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*. [Online] 18 (4), 497–511.

Rivas, M. A. et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*. [Online] 43 (11), 1066–1073.

Rivera, M. et al. (2012) Depressive disorder moderates the effect of the FTO gene on body mass index. *Molecular Psychiatry*. [Online] 17 (6), 604–611.

Schifano, E. D. et al. (2013) Genome-wide Association Analysis for Multiple Continuous Secondary Phenotypes. *The American Journal of Human Genetics*. [Online] 92 (5), 744–759.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. [Online] 511 (7510), 421–427.

Segura, V. et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*. [Online] 44 (7), 825–830.

Selzam, S. et al. (2016) Predicting educational achievement from DNA. *Molecular Psychiatry*. [Online] [online]. Available from: http://www.nature.com/mp/journal/vaop/ncurrent/full/mp2016107a.html (Accessed 5 September 2016).

Servin, B. & Stephens, M. (2007) Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLOS Genet*. [Online] 3 (7), e114.

Shim, H. et al. (2015) A Multivariate Genome-Wide Association Analysis of 10 LDL Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLoS ONE*. [Online] 10 (4), e0120758.

Shin, S.-Y. et al. (2014) An atlas of genetic influences on human blood metabolites. Nature genetics. [Online] 46 (6), 543–550.

Sidak, Z. (1968) On Multivariate Normal Probabilities of Rectangles: Their Dependence on Correlations. *The Annals of Mathematical Statistics*. [Online] 39 (5), 1425–1434.

Sidak, Z. (1971) On Probabilities of Rectangles in Multivariate Student Distributions: Their Dependence on Correlations. *The Annals of Mathematical Statistics*. [Online] 42 (1), 169–175.

van der Sluis, S. et al. (2013) TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. *PLoS Genet*. [Online] 9 (1), e1003235.

Smith, D. J. et al. (2016) Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Molecular Psychiatry*. [Online] 21 (6), 749–757.

Smith, D. J. et al. (2013) Prevalence and Characteristics of Probable Major Depression and Bipolar Disorder within UK Biobank: Cross-Sectional Study of 172,751 Participants. *PLOS ONE*. [Online] 8 (11), e75362.

Solovieff, N. et al. (2013) Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*. [Online] 14 (7), 483–495.

Staley, J. R. et al. (2016) PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*. [Online] btw373.

Stephens, M. (2013) A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE*. [Online] 8 (7), e65245.

Stephens, M. & Balding, D. J. (2009) Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*. [Online] 10 (10), 681–690.

Sudlow, C. et al. (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. [Online] 12 (3), e1001779.

Sullivan, P. F. et al. (2000) Genetic Epidemiology of Major Depression: Review and Meta-Analysis. *American Journal of Psychiatry*. [Online] 157 (10), 1552–1562.

Taal, H. R. et al. (2012) Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nature Genetics*. [Online] 44 (5), 532–538.

Teslovich, T. M. et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. [Online] 466 (7307), 707–713.

The Early Growth Genetics (EGG) Consortium (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature Genetics*. [Online] 44 (5), 526–531.

The International Consortium for Blood Pressure (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. [Online] 478 (7367), 103–109.

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*. [Online] 467 (7311), 52–58.

The Tobacco and Genetics Consortium (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*. [Online] 42 (5), 441–447.

Valk, R. J. P. van der et al. (2015) A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human Molecular Genetics*. [Online] 24 (4), 1155–1168.

Vassos, E. et al. (2016) An Examination of Polygenic Score Risk Prediction in Individuals with First Episode Psychosis. *Biological Psychiatry*. [Online] 0 (0), . [online]. Available from: /article/S0006-3223(16)32664-6/abstract (Accessed 29 September 2016).

Vattikuti, S. et al. (2012) Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. *PLoS Genet*. [Online] 8 (3), e1002637.

Visscher, P. M. et al. (2012) Five Years of GWAS Discovery. *The American Journal of Human Genetics*. [Online] 90 (1), 7–24.

Visscher, P. M. & Yang, J. (2016) A plethora of pleiotropy across complex traits. *Nature Genetics*. [Online] 48 (7), 707–708.

Whiteford, H. A. et al. (2013) Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*. [Online] 382 (9904), 1575–1586.

Willer, C. J. et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*. [Online] 40 (2), 161–169.

Wood, A. R. et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*. [Online] 46 (11), 1173–1186.

Yang, J. et al. (2012) FTO genotype is associated with phenotypic variability of body mass index. *Nature*. [Online] 490 (7419), 267–272.

Yang, J. et al. (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*. [Online] 88 (1), 76–82.

Zhang, Y. et al. (2014) Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*. [Online] 96309–325.

Zhou, X. et al. (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genet*. [Online] 9 (2), e1003264.

Zhou, X. & Stephens, M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*. [Online] 11 (4), 407–409.

Zhou, X. & Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. [Online] 44 (7), 821–824.

Zhu, X. et al. (2015) Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. *The American Journal of Human Genetics*. [Online] 96 (1), 21–36.

Zhu, X. & Stephens, M. (2016) *Bayesian large-scale multiple regression with summary statistics from genome-wide association studies*. [online]. Available from: http://biorxiv.org/lookup/doi/10.1101/042457 (Accessed 18 September 2016). [online]. Available from: http://biorxiv.org/lookup/doi/10.1101/042457 (Accessed 18 September 2016).